ARI Research Note 92-74

Selection and Classification Tests for Critical Military Occupational Specialties

Scott H. Oppler, Norman G. Peterson, Deborah L. Whetzel, Diane Steele, Ruth A. Childs, Randolph K. Park, and Rodney L. Rosse

American Institutes for Research

John F. Rehling, Thomas M. Brantner, and William F. Kieckhaefer

RGI, Inc.

Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Research Division
Zita M. Simutis, Director

August 1992

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

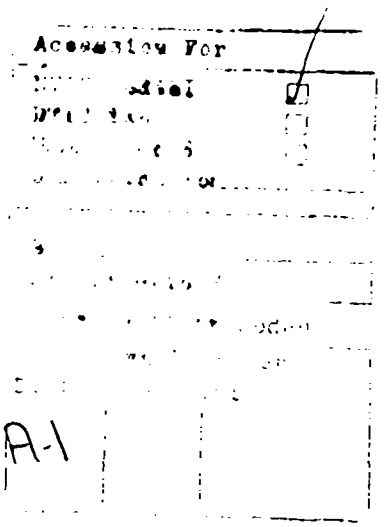
MICHAEL D. SHALER
COL, AR
Commanding

Research accomplished under contract
for the Department of the Army

American Institutes for Research

Technical review by

Clinton B. Walker



NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1992, August	Final Sep 89 - Sep 91

4. TITLE AND SUBTITLE

Selection and Classification Tests for Critical Military Occupational Specialties

5. FUNDING NUMBERS

MDA903-89-C-0266
63007A
792
2207
C2

6. AUTHOR(S)

Oppler, Scott H.; Peterson, Norman G.; Whetzel, Deborah L.; Steele, Diane; Childs, Ruth A.; Park, Randolph K.;
(Continued)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

American Institutes for Research
3333 K Street, NW, Suite 300
Washington, DC 20007

8. PERFORMING ORGANIZATION REPORT NUMBER

--

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-R
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ARI Research Note 92-74

11. SUPPLEMENTARY NOTES

Contracting Officer's Representative, Michael G. Rumsey.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release;
distribution is unlimited.

12b. DISTRIBUTION CODE

--

13. ABSTRACT *(Maximum 200 words)*

This report presents research on issues affecting the readiness of new Army personnel tests for implementation. These tests are from the battery developed under the Army's Project A to supplement the Armed Services Vocational Aptitude Battery (ASVAB). Specific tests were chosen for this research because they are being considered by the Department of Defense for implementation in preenlistment selection and classification. For this research, in 1990 new recruits at Fort Knox and Fort Benning took computerized tests of spatial, perceptual, and psychomotor abilities. Three possible sources of extraneous variance in the test scores were investigated: order effects, practice effects, and effects of changes in test instructions. In addition, pencil and paper versions of new biodata and spatial items were tested. As part of the work to counteract falsification on self-report instruments, forms of the Assessment of Background and Life Experiences (ABLE) were administered under various instructional sets. Finally, scores for two of the computerized tests were compared with archival scores on pencil-and-paper versions from
(Continued)

14. SUBJECT TERMS

Psychomotor tests
Able
Project A spatial tests

15. NUMBER OF PAGES

130

16. PRICE CODE

--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

ARI Research Note 92-74

6. AUTHOR(S) (Continued)

Rosse, Rodney L. (AIR); Rehling, John F.; Brantner, Thomas M.; and Kieckhaefer, William F. (RGI, Inc.)

13. ABSTRACT (Continued)

Project A to determine possible effects of medium of administration on test performance. Approaches to dealing with the sources of variance other than ability are discussed, and needs for follow-on research are identified.

SELECTION AND CLASSIFICATION TESTS FOR CRITICAL MILITARY OCCUPATIONAL SPECIALTIES

EXECUTIVE SUMMARY

Requirement:

For this report, researchers collected and analyzed data to investigate psychometric and mode of administration issues for Project A/Career Force measures that are candidates for supplementing the traditional Armed Services Vocational Aptitude Battery (ASVAB).

Procedure:

Recruits at Fort Benning and Fort Knox were tested on computerized spatial, psychomotor, perceptual speed, and temperament/biodata measures from May 1990 to August 1990. Test score and latency data were collected to examine the effects of changing the clarity of instructions on tests, changing the order of items in the temperament/biodata measure, and using different modes of administration for the spatial and temperament/biodata measures. The effects of practice and test order on psychomotor and perceptual speed tests were also evaluated. Data from Project A were included in some analyses to make appropriate comparisons between test versions. New forms of the Assessment of Background and Life Experiences (ABLE) and a spatial item pool were developed and administered. Additionally, ABLE was administered under various instructional sets.

Findings:

This research produced several major findings, including the following: changes to Project A/Enhanced Computer Administered Tests (ECAT) instructions did not produce noticeable improvements; changing the order of items in the temperament/biodata instrument made little difference in psychomotor characteristics of the instruments' scales; changing the mode of administration from paper-and-pencil to computer administration positively affects the characteristics of ABLE scale scores without affecting mean scale scores and produces lower mean scores on spatial tests without affecting other psychometric characteristics; practice substantially improves performance on psychomotor and perceptual speed tests; and test order makes little difference in terms of the mean scores achieved and the reliability of measurement.

Utilization of Findings:

This report will be used by the Deputy Chief of Staff for Personnel and the Assistant Secretary of Defense (Force Management and Personnel) to further document the utility of using Project A/Career Force measures for enlisted selection and classification and to direct future research efforts on these measures.

SELECTION AND CLASSIFICATION TESTS FOR CRITICAL MILITARY OCCUPATIONAL SPECIALTIES

CONTENTS

CONTENTS (Continued)

LIST OF TABLES

CONTENTS (Continued)

CONTENTS (Continued)

CONTENTS (Continued)

LIST OF FIGURES

CONTENTS (Continued)

SELECTION AND CLASSIFICATION TESTS FOR CRITICAL MILITARY OCCUPATIONAL SPECIALTIES

CHAPTER 1

INTRODUCTION

This report describes activities carried out as part of the Selection and Classification Tests for Critical Military Occupational Specialties (CMOS) project. As initially conceived, the project had two major purposes: first, to install, manage, evaluate, and report on Army testbeds of newly developed personnel tests; and second, to upgrade the testbeds by improving or adding to the personnel tests and performance criteria that they encompassed. Each testbed was a particular implementation of newly developed Army personnel tests that were administered to soldiers during their entry-level training under fairly realistic operational conditions.

As the project started, major and unforeseen events occurred that shifted emphasis from the first to the second purpose. In Europe, international events precluded establishing the testbeds in the active forces. In the continental United States, the Army withdrew testbeds established at several training sites to make way for a testbed sponsored by the Department of Defense (DoD). The DoD testbed overlapped with the Army testbeds, both in terms of the personnel tests that were to be evaluated and the training posts at which data were to be collected.

As a result of these changes, the CMOS project has concentrated on an investigation of psychometric and mode of administration issues for Project A/Career Force measures that are candidates for supplementing the ASVAB. The research focused on the six Army ability tests that were included as part of the nine tests in the DoD battery and on a temperament/biodata inventory that was being considered for use as part of another DoD testing initiative. In the chapters that follow we report on research addressing a number of issues that surround the possible use of these tests in an operational setting. First, however, we provide background information on the development and initial validation of the tests, discuss pertinent research carried out in the Army testbeds, and describe other ongoing research efforts that helped shape the activities of this project.

Project A Origins

The tests that are the objects of the research reported here have their origins in earlier research sponsored by the Army Research Institute (ARI) known as Project A. The earlier research had as its goal the generation of "...the criterion variables, predictor measures, analytic methods, and validation data that are necessary for developing an enhanced selection and classification system for all entry-level positions in the United States Army (Campbell, 1990a, pp. 232)." Two important subgoals

of Project A were the validation of the Armed Services Vocational Aptitude Battery (ASVAB) against job performance criteria and the development and validation of new predictor measures against those same criteria.

The results of Project A were successful and they are documented in several ARI Technical Reports (Campbell, 1988; Campbell & Zook, 1991; Peterson, 1987) as well as a special issue of Personnel Psychology (Campbell, 1990b) devoted to the project. For our purposes, we note that Figure 1-1 shows the names of the 19 new predictor measures that were developed and validated as part of Project A.

Cognitive Paper-and-Pencil Tests	Reasoning Test*
	Assembling Objects Test*
	Orientation Test*
	Object Rotation Test
	Maze Test
	Map Test
Computer-Administered Tests	Target Tracking Test 1*
	Target Tracking Test 2*
	Target Identification Test*
	Reaction Time 1
	Reaction Time 2
	Memory Test
	Number Memory Test
	Cannon Shoot Test
	Perceptual Speed and Accuracy Test
	Target Shoot Test
Paper-and-Pencil Inventories	Assessment of Background and Life Experiences (ABLE)*
	Army Vocational Interest Career Examination (AVOICE)
	Job Orientation Blank (JOB)

Figure 1-1. Array of measures included in Project A Predictor Batteries (*tests used in the current research).

Of the measures shown in Figure 1-1, seven were used for this project. These include three of the cognitive paper-and-pencil tests, three of the computer-administered tests, and one of the paper-and-pencil inventories.

The three cognitive paper-and-pencil tests were all developed to measure various aspects of spatial ability. The Reasoning Test was developed to measure the ability to generate hypotheses about principles governing relationships among several objects. In this test, the examinee is presented with a series of four figures and asked to identify from among five possible answers the one figure that should appear next in the series. The Assembling Objects test was developed to measure the ability to mentally manipulate components of two- and three-dimensional figures into other arrangements, specifically the rotation of such figures. In this test, the examinee is presented with the components or parts of an object and asked to select, from among four alternatives, the one object that depicts the parts assembled correctly. The Orientation test was developed to measure the ability to maintain one's bearings with respect to landmarks. Each item stem shows a framed picture. The picture, which is tilted, is said to be fixed in place, but the frame is said to rotate around it. Scribed on the frame at the side or top of the picture is a small pattern. Examinees are asked to rotate the frame mentally until the small pattern is at the foot of the picture, and then to pick from multiple choices the one that shows how the pattern will look in this new position.

Two of the three computer-administered tests involve tracking tasks. Target Tracking Test 1 was developed to measure psychomotor precision. This test measures ability to make muscular movements necessary to adjust or position a control mechanism, in this case to track a target moving on a predictable path at a predictable speed. In the actual task, the examinee uses a joystick to keep a crosshair centered in the middle of a box moving along a visible path at a constant speed on the computer screen. Target Tracking Test 2 uses the same task, except the crosshair is controlled by two sliding resistors, one for each hand. It was developed to measure multilimb coordination--the ability to coordinate the movement of two or more limbs while the trunk is at rest (seated or standing). The third computer-administered test was developed as a measure of perceptual speed and accuracy, the ability to perceive visual information quickly and accurately and to perform simple tasks with the information (e.g., make comparisons). The Target Identification Test presents the examinee with a "target object" near the top of the computer screen and three other objects in a row near the bottom of the screen. The examinee's task is to quickly identify which of the three objects at the bottom of the screen matches the target object and to press the button corresponding to that object.

The Assessment of Background and Life Experiences (ABLE) instrument is an untimed, paper-and-pencil inventory that was developed to measure eleven substantive scales that tapped seven distinct temperament/biodata constructs. The ABLE includes four response validity scales. Figure 1-2 presents the names of the seven constructs measured by the eleven substantive scales and gives the names of the four response validity scales. The names are fairly descriptive of the content of the constructs and scales, but more detail can be found in Peterson (1987) and Peterson, Hough, Dunnette, Rosse, Houston, and Toquam (1990).

Construct	Scale
Adjustment	Emotional Stability
Dependability	Nondelinquency
	Traditional Values
	Conscientiousness
Achievement	Work Orientation
	Self-Esteem
Physical Condition	Physical Condition
Leadership (Potency)	Dominance
	Energy Level
Locus of Control	Internal Control
Agreeableness/Likability	Cooperativeness
Response Validity	Non-Random Response
	Fake Good (Social Desirability)
	Fake Bad
	Self-Knowledge

Figure 1-2. Scales (by construct) included in the Assessment of Background and Life Experiences (ABLE)

Project A job performance criterion measures were made up of job knowledge tests, hands-on tests, and supervisory/peer ratings (Campbell, Ford, Rumsey, Pulakos, Borman, Felker, De Vera, and Riegelhaupt, 1990). They were highly reliable and measured important aspects of soldiers' jobs, but they did not measure fairly narrow, dynamic parts of soldier's jobs--like gunnery performance. Project A validation results showed that the ASVAB

does a very good job of predicting MOS-specific and Army-wide technical performance on soldiers' jobs, but does less well at predicting the motivational and disciplinary aspects of the job (McHenry, Toquam, Hough, and Ashworth, 1990). The new predictors developed as part of Project A showed small amounts of validity over and above that provided by the ASVAB (incremental validity) for the technical performance criteria, but substantially greater amounts of incremental validity for the motivational and disciplinary criteria.

The validation results reported for the new Project A tests (e.g., McHenry, et al., 1990) were based on composites of the new Project A tests. That is, scores from several tests were combined to produce a composite score. Therefore, the validity of individual Project A tests was unknown. Busciglio, Silva, and Walker (1990) investigated the validity of the six Project A spatial tests, ten scores from the computer-administered tests, and the ASVAB subtests. (The ABLE was not included in their work.) They used the data collected as part of Project A to determine the validity of each test and the extent to which the Project A tests added to the validity of the ASVAB. They concluded that the new tests predicted from 4% to 13% more criterion variance than when the ASVAB was used alone. They also found that both the paper-and-pencil spatial tests and the computer-administered tests of perceptual speed and psychomotor ability added to the validity of ASVAB in significant ways.

The result of the Project A research was the development and validation of a set of new tests against job performance criteria. However, only one form was available for each test and the validation was against an inclusive, but fairly general set of criterion measures. (We note that a new project, called Building the Career Force, has been undertaken to build on the results from Project A for predicting later career performance.) If the tests were to be used in an operational setting, it would be extremely useful to have additional forms of the tests developed and to conduct research to determine their validity for specific, critical Army tasks, in particular, gunnery. We now turn to a brief description of research conducted in Army testbeds that was primarily aimed at this last goal, investigating the validity of the new tests against measures of gunnery performance.

Army Testbeds

Encouraged by the Project A results, Army decision makers ordered the implementation of some of the Project A tests in the Training and Doctrine Command's (TRADOC) Skills Selection and Sustainment (S3) Program. This program was ordered by General Maxwell Thurman and called for, among other things, the testing of leadership potential and spatial and psychomotor abilities for selection and fast tracking into gunnery occupational specialties

(Walker, 1989). The Army Research Institute assisted the field sites in implementing the testing aspects of the program. ARI provided several types of support which included: supplying computer-based test stations, software and training necessary for administering Target Tracking Tests 1 and 2; manuals and material for administering the ABLE inventory and the Maze and Orientation tests; suggested scoring schemes for combining test scores; and ongoing technical support. During the course of this program, ARI also converted the ABLE to a computer-administered format. The testbeds were put into place at Forts Benning, Knox, Bliss, and Sill, though testing never got underway at Ft. Sill.

Over 25,000 soldiers completed the S3 Program's test battery during their training. The soldiers completed the battery with the understanding that their test scores counted and would be recorded in their records. (In Project A, the soldiers had been informed that test results would not be entered into their records.) Analysis of the relationship of S3 Program test scores to training performance, as measured on realistic simulators and by time taken to attain qualifying levels of performance, indicated that the new tests worked well for identifying students with aptitudes to be good tank and anti-tank gunners. The tests did not do very well at identifying good anti-aircraft specialists. ARI researchers concluded that the lack of validity for the anti-aircraft specialty was most likely due to the different mix of abilities that appear to be required for this job when compared to tank and anti-tank gunnery specialties.

Busciglio, Silva and Walker (1990) report testbed results in detail for anti-tank gunners (MOS 11H) and M1 Armor Crewmen (MOS 19K). Both tracking tests significantly predicted simulator performance for anti-tank gunners (correlations were .28 and .33 for one-hand and two-hand tracking, respectively). The Target Tracking Test 2 (two-hand tracking) and Maze test showed incremental validity, predicting an additional 9.6 percent of criterion variance over that achieved by ASVAB subtests. Results for armor crewmen showed that all of the Project A predictors were significantly related to simulator performance at levels comparable to ASVAB subtests (validity coefficients ranged from .35 to .43 for the four Project A tests and ~.06 to .42 for ASVAB subtests), and that Target Tracking Test 2 and the Maze test again added incremental validity, accounting for an additional 10.5% of criterion variance.

These testbed results seemed to indicate that use of the Project A tests "...can markedly improve the quality of Army enlisted personnel (Busciglio, et al., 1990, pp. 12)." However, as they also noted, difficult practical problems were encountered by the S3 Program personnel when trying to use test scores to assign soldiers. One solution to these problems, they noted, would be to administer the new tests during pre-enlistment processing of civilian applicants, rather than waiting to test

soldiers after they have enlisted and reached their training post.

Enhanced Computer Assisted Testing (ECAT)

The Department of Defense has instituted a program to evaluate the potential benefit of adding new, computer-administered tests to the ASVAB for purposes of pre-enlistment processing. This program, called Enhanced Computer Assisted Testing (ECAT), included tests chosen from a set of tests nominated by the Services. Although some of the nominated tests had been administered in paper-and-pencil formats, all of the tests were converted to computer-administration on the Hewlett-Packard (HP) Integrated PC for this project. (Appendix A describes the computer programming and apparatus developed for the ECAT project and used to collect data on the ECAT battery.) The Navy took the lead in preparing the battery, after its content had been selected by all of the Services. Each Service made several MOS available for ECAT testing, which has been conducted by RGI, Inc. Both test and criterion data have now been collected, and the Navy will soon report on the results of their analyses of these data.

The institution of data collection for the DoD-sponsored ECAT battery in the Services essentially displaced the Army's own S3 program battery. Data were to be collected at many of the same sites, and it was not feasible to collect data for both batteries. However, as noted earlier, the nine-test ECAT battery contained six of the Army's Project A tests, including all but one of the spatial and psychomotor/perceptual speed tests included in the S3 Program (the Maze test was not included in the ECAT battery). Also, the goal of ECAT was the same as the Army's goal: to evaluate the new tests for possible inclusion in pre-enlistment processing.

As we noted at the beginning of this chapter, CMOS research activities concentrated on the six Project A tests included in the ECAT battery since these seemed most likely to be implemented, and the ECAT administration methodology (the computer hardware and software) was available for our use. As a reminder, three of these six tests were the spatial tests: Assembling Objects, Orientation, and Spatial Reasoning. (The Spatial Reasoning test was named Reasoning Test and, sometimes, Figural Reasoning during Project A, but was called Spatial Reasoning in the ECAT project. We have adopted the latter convention and also have dropped "Test" from the names of the other two tests.) Three computer-administered tests were included, the two tracking tests--Target Tracking Test 1 (called One-Hand Tracking for ECAT and here) and Target Tracking Test 2 (Two-Hand Tracking) and the Target Identification test.

ABLE and the Applicant Screening Profile (ASP)

Although the ABLE was not included in the ECAT battery, DoD has seriously investigated the use of temperament and biodata instruments for use in pre-enlistment processing. This serious consideration, with a strong possibility for future operational use of temperament/biodata measures, provided a strong rationale for further research with the ABLE. DoD recently sponsored the development of the Applicant Screening Profile (ASP), an instrument which is a combination of some of the ABLE scales and biodata items developed and validated by the Navy. Plans were made to collect data on the ASP in an operational setting and to use it in some fashion to screen civilian applicants. These plans were not carried through as quickly as anticipated, but the ASP remains a strong possibility for implementation in the operational setting.

For these reasons we included the ABLE in CMOS research. We used the ABLE in a computer-administered format since ARI had already administered it this way in its testbed, and it seemed likely to be so used in actual operation.

Summary

Project A developed and validated a new set of personnel tests for Army enlisted ranks. Subsequent Army testbed research provided further, more focused evidence that the new spatial and computer-administered tests hold considerable promise for improving the quality of gunnery performance by Army enlisted personnel. Department of Defense initiatives pointed to the use of these types of tests in computer-assisted format for pre-enlistment processing. All of these factors pointed strongly to the evaluation of issues concerned with ongoing, operational use of the new tests in paper-and-pencil and computer-administered formats.

One such issue is that use of the new tests in pre-enlistment processing would lead to coaching and practicing for taking the tests. The abilities tested by the ASVAB, particularly by the Armed Forces Qualifying Test (a composite of the verbal and mathematical subtests on the ASVAB), are heavily coached and practiced as a matter of course in the American educational system. However, the kinds of abilities and types of items on most of the new Project A tests are much less familiar. For example, two-hand tracking and identification of targets are not taught widely in the United States. As a result, scores on the new tests could reflect an unknown and unstable mix of effects attributable to ability, coaching, practice, sequence of tests, and even test instructions. It is important to determine the level and nature of these effects and, if unwanted effects exist, to develop procedure for controlling or eliminating them.

Another important issue concerns testing format. Several of the new tests were originally developed and validated in paper-and-pencil formats, but have since been converted for computer administration. Research must be conducted to determine that the construct being measured is the same across formats and to determine the effect of format on score levels and other psychometric properties. A related issue concerns the effect of order of item administration for temperament/biodata instruments like the ABLE. Computer administration opens up possibilities for presenting such items in different orders, perhaps to help prevent memorization of inventory content by applicants.

This report documents the activities undertaken to investigate these and a related set of testing issues. Two major types of activities were completed. The first was the design and conduct of experiments to investigate several of the issues alluded to above. The second was the collection of data necessary to complete other investigations planned by ARI. The balance of this report primarily describes the first type of activity. Appendix B includes a complete description of the data collected as part of the CMOS project, including those data to be analyzed outside this project.

Chapter 2 describes an experiment conducted to determine the effects of modifying the instructions of the spatial and psychomotor/perceptual speed tests, to make them as consistent and user-friendly as possible, and an experiment to examine the effects of varying the order in which the ABLE items are administered. Chapter 3 examines the effects of mode of administration (computer versus paper-and-pencil) for the spatial tests and the ABLE. Chapter 4 addresses the effect of practice on scores obtained on the psychomotor and perceptual speed tests. In Chapter 5 we examine the effect of order of administration of the psychomotor and perceptual speed tests. Chapter 6 describes the collection of data for the evaluation of new items developed for the spatial tests, including suggestions for analysis of the collected data. Chapter 7 summarizes the results of the prior chapters and suggests next steps.

CHAPTER 2

CLARITY OF TEST INSTRUCTIONS AND ORDER OF ABLE ITEMS

In this chapter, two experiments are described. One experiment investigated the effects of modified instructions, and the other investigated the effects of order of item administration. These two, different experiments are described in one chapter because the data for the experiments were collected on the same samples and many of the analyses are similar.

Tests used in the experiment comparing old and new instructions were the Hewlett-Packard (HP) computer versions of the six Project A tests (three spatial, three psychomotor/ perceptual speed) used for ECAT testing. The spatial ability tests were Assembling Objects, Orientation, and Spatial Reasoning. The psychomotor/perceptual speed tests were One- and Two-Hand Tracking and Target Identification.

The computerized version of the Assessment of Background and Life Experiences (ABLE), originally programmed to be administered on the Seequa as part of the Army testbeds (see Chapter 1), was also transported to the HP to conduct an experiment on order of administration of the items on the ABLE. Many of the concerns with the instructions for the spatial and psychomotor tests (see immediately below) did not apply to the ABLE, primarily because the instructions for the ABLE are extremely simple and brief. There was, however, some concern that ABLE scale scores could be affected by the order in which the items are administered on the computer. Therefore, the ABLE instructions were not changed but the ABLE was administered with two item orders: the original and a reversed order, i.e., the last item appeared first, the second to the last item appeared second, and so on.

The test instructions became a target of research for several reasons. The first was to ensure that scores on tests of particular attributes were influenced as little as possible by the examinees' other attributes. To the extent that the reading level of the nations's youth has decreased in recent years (Johnston & Packer, 1987), the clarity of instructions used for tests that measure constructs other than reading ability has become an issue. In the present instance, we wanted to ensure that scores on the psychomotor, perceptual speed, and spatial tests were not strongly influenced by examinee reading ability. By simplifying the instructions for such tests, we hoped to reduce any differences in test scores that might be due to differences in understanding of instructions.

Second, selection tests like these, which were developed when the Services could recruit good readers, need to be usable

when good readers are not as available for enlistment. Indeed, the utility of the new Project A tests would probably be greater at such times; soldiers' success on the job would depend more on their other aptitudes than on reading ability.

Third, despite the careful development of the Project A tests and their instructions (see below), the reading times for the computerized tests and the formatting of their instructional screens raised concerns that the instructions were still not easy enough to read. For these reasons, AIR's Document Design Center reviewed and revised the instructions for the six ECAT tests. We then investigated the impact of the revised instructions on instruction reading time, test-taking time, test scores, ratings of instruction clarity, and the relation of test scores to independent measures of cognitive ability.

Background on Development of Instructions

Significant effort previously went into the development of the Project A tests and instructions. After the tests were initially developed, they were pilot tested (up to three times) and field tested (once) using an iterative procedure, in which feedback obtained during each stage of pilot testing was incorporated in subsequent changes to the tests and instructions. During these iterations, the tests were administered to (and feedback was received from) as many as 500 examinees per test (Peterson, 1987). Feedback was obtained by asking examinees to describe their general reactions to the tests, including the clarity and completeness of instructions, the perceived difficulty of the tests, and (for the psychomotor/perceptual speed tests) the ease of use of the response pedestal.

As a result of field testing, additional changes were made. In particular, the instructions for the psychomotor/perceptual speed tests were shortened considerably, and a standard outline was followed when preparing the revised instructions. The outline consisted of: 1) the test name, 2) a one-sentence description of the purpose of the test, 3) step-by-step instructions, 4) a preliminary practice item, 5) a brief restatement of the test instructions, 6) two or three additional practice items, and 7) instructions to call the test administrator if there were further questions about the test. The tests were then administered on a large scale to the Project A concurrent validation (CV) sample. Following CV testing, only minor changes were made to the test instructions prior to administration of the tests to the Project A longitudinal validation (LV) sample.

Finally, as indicated above, computerized versions of three of the Project A paper-and-pencil spatial tests were developed for use in the ECAT battery, and three of the psychomotor/ perceptual speed tests were transported for use with the ECAT

software and hardware (i.e., the HP). Additionally, a computerized version of the ABLE was developed and transported to the HP for the present research.

Generally, the instructions of the HP versions of these tests were intended to be the same as those for the Project A (Seequa or paper-and-pencil) versions. However, in some cases there were differences. For example, for the Orientation test, a dynamic (i.e., moving) sample item demonstrating the relation between the item stem and the correct option was incorporated into the computerized instructions. Other differences included the addition of instructions telling the examinee how to use the response pedestal to enter the intended response for those tests that had previously been administered via paper-and-pencil (i.e., the spatial tests).

Sample

The analyses described below are based on a total of 437 examinees tested at the Fort Benning, Georgia, Reception Battalion. Of these 437 examinees, 207 were regular Army recruits with no prior enlistment experience. Of these recruits, 205 were assigned to the 11X Infantry Military Occupational Specialty; one was assigned to 91A; and the other was unknown. The remaining 230 examinees consisted of recruits for the Army Reserves and the National Guard.

Variables and Measures

Scoring of tests and collection of other data. As noted above, all seven instruments were administered via computer. Two measures were collected for the six original ECAT tests: time to complete instructions and time to complete the test. These "time to complete" scores were automatically recorded by the computer. Time to complete instructions included the time from the appearance of the first screen of instructions to the time of appearance of the first test item. Time to take the test included the time from the beginning of the first test item to the completion of the last test item. After completing each test, examinees rated how difficult it was to understand the instructions. The options were "very hard," "somewhat hard," and "not hard at all."

For the three spatial tests and the ABLE, two pieces of information were recorded by the computer for each item response: time to respond to the item and response option chosen. For the ABLE, times to complete each item were summed to obtain the time to complete the full ABLE. Application of appropriate scoring keys provided number correct scores for the spatial tests and scale scores for the ABLE.

ABLE scores for two examinees were treated as missing because of their scores on the ABLE's Non-Random Response scale. The intent of this scale (Campbell, 1987) is to detect respondents who cannot or are not reading the questions, and instead are randomly responding. Also, responses to five items from the ABLE were treated as missing for all examinees because of text errors in the computer-administered versions. (Exactly one item was affected for each of the following five scales: Physical Condition, Dominance, Nondelinquency, Traditional Values, and Internal Control. Scores for these scales were then adjusted using missing data scoring rules developed in Project A, so that the maximum possible score on each scale was not affected by the missing item score.)

Scoring for One-Hand and Two-Hand Tracking tests was identical. As we described in Chapter 1, examinees on these tests are shown (on the computer monitor) a path consisting of vertical and horizontal line segments, at the beginning of which is a target box with a crosshair in the center. The target moves along the path at a constant speed and the examinee's task is to keep the crosshair centered within the target at all times. A joystick is used to control crosshair movement for One-Hand Tracking while two sliding resistors are used for the same purpose for Two-Hand Tracking. Each test consists of 18 scored test items plus six unscored practice items (the first three of which are administered as part of the self-paced instructions). The score for each item is the natural logarithm of the average root mean square distance from the center of the crosshair to the center of the target as computed every 50 milliseconds that the target is on the screen (total time on screen = 9.99 seconds). The score for the total test is obtained by computing the mean of the 18 item scores. This score is referred to as the mean log distance score.

For each item on the Target Identification test, examinees are presented (on the computer monitor) a target object and three stimulus objects. The objects are pictures of military vehicles or aircraft. The examinee must determine which of the three stimulus objects is the same as the target object and then press a button on the response pedestal corresponding to that choice. An item is presented only when examinees have placed their hands in the "ready" position, i.e., depressed four buttons labelled "home," thus insuring that their hands are always in the same position prior to each item. After the item appears, releasing the home buttons makes the target vanish. There are 36 scored items on Target Identification plus three unscored practice items (all of which are administered as part of the self-paced instructions). Three scores are computed for each item: response accuracy, decision time, and movement time. Response accuracy simply refers to whether the correct stimulus object (i.e., that matching the target object) was identified. Decision time refers to the time between the onset of the item and the

point at which the examinee removes his/her hand from the "home" buttons on the response pedestal. This interval reflects the time required to process the information to determine the correct response. Movement time refers to time between the release of the home buttons and the pressing of a response key.

Three test scores, corresponding to the three types of item scores, are derived for each examinee. The first of these, percent correct, is simply the percentage of items on which the examinee responded correctly. The other two scores are the "clipped mean decision time" and the "median movement time." These scores were developed based on analyses of data collected from the Project A longitudinal validation sample (Campbell & Zook, 1991), during which several optional scoring schemes were evaluated. The clipped mean decision time and median movement time scores are computed using data from only those items that were answered correctly. To compute clipped mean decision time, the correctly answered items are first divided into two sets: "difficult" and "easy." The difficult set includes items in which the three stimulus objects are the same type as the target object (e.g., three different varieties of tanks), whereas the easy set includes items in which only two of the three stimulus objects are of the same type (e.g., two tanks and one truck). For each set of items, the mean decision time is computed following the elimination of both the fastest and slowest decision times. The final score is the average of these two clipped means. To compute the median movement time score, the distinction between difficult and easy items is ignored. Instead, this score is simply the median of the movement times associated with correctly answered items. Note that decision time and median time scores are computed only for examinees responding correctly to at least 33 percent of the items.

Instructions and order of administration. Revised instructions were developed for each test. In some cases, attempts to improve readability were made by reducing the amount of information provided on a single instruction screen. That is, information which had originally been presented on a single screen in the original instructions was often spread out over two or three screens in the revised instructions. All instructions were also reviewed for idiosyncratic language and inconsistent language usage. For example, the instruction "Hit the button" was changed to "Press the button."

The seven tests were administered in a fixed sequence beginning with the ABLE, followed by the spatial tests (in order of Assembling Objects, Orientation, and Spatial Reasoning) and the psychomotor/perceptual speed tests (in order of One-Hand Tracking, Two-Hand Tracking, and Target Identification). One sample of 219 examinees was tested with the newly revised instructions, and another sample of 218 examinees was tested with the Project A/ECAT instructions. For the ABLE, one examinee in

the Project A/ECAT instructions group received the items in reversed order, but all others in this group received the original order of items. The examinees in the new instruction group all received the ABLE items in reversed order.

Analyses

Instructions experiment. Means and standard deviations were computed for instruction reading time, test-taking time, test scores, and clarity of instructions for tests with old and new instructions. Student t-tests (Hays, 1963) were computed to test for significant mean differences between tests with the old and new instructions. Similarly, comparisons were made of the internal consistency reliabilities for tests with old and new instructions. Finally, correlations of AFQT with instruction time and test scores for the two versions were compared.

ABLE item order experiment. Means and standard deviations were computed for time to complete the inventory and scale scores for the original and reversed orders of presentation. Student t-tests were computed to test for significant mean score differences and F-tests (Hays, 1963) were computed to test for significant differences in scale score variances. Internal consistency reliabilities for the scales on the two versions were computed and compared. Correlations of ABLE scale scores with AFQT scores were not computed since results from very large samples indicate that these correlations are very nearly zero (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990).

Results: Instructions Experiment

Instruction Times. The average times for reading the old and new instructions for the spatial and psychomotor/perceptual speed tests are shown in Table 2-1. Student t-tests indicated no statistically significant differences ($p>.05$) between the mean time to read the old versus new instructions.

Test-Taking Times. The average test-taking times (that is, the time required to take all of the items) for examinees who received the old or new instructions are reported in Table 2-2. Student t-tests indicated no significant differences ($p>.05$) in the average test-taking times between the two sets of examinees for any of the tests. It should be noted that the items on the One- and Two-Hand Tracking tests are administered at approximately the same pace for all examinees; however, variations in test-taking time can occur on these tests when examinees pause the test by pressing the HELP button on the response pedestal.

Test Scores. Table 2-3 presents the means and standard deviations of test scores on the spatial and psychomotor/perceptual speed tests. Note that the scores for the three spatial tests are number correct scores; scores for the two tracking tests are mean log distance scores; and scores for the Target Identification test are percent correct, mean decision time, and median movement time. Procedures used to compute the scores for the two tracking tests and Target Identification are described earlier in this chapter. Student t-tests comparing the means reported in Table 2-3 indicated no significant differences ($p>.05$) between examinees who received the old or new instructions.

Reliabilities. Table 2-4 shows the reliability estimates of the spatial and psychomotor/perceptual speed tests for examinees who received the old or new instructions. The reliabilities for the spatial tests are coefficients alpha, and the reliabilities for the psychomotor/perceptual speed tests are split-half reliabilities (based on the correlations between odd and even items). The split-half reliabilities have been corrected using the Spearman-Brown formula (Cascio, 1987). The reliabilities of the tests appear to have been unchanged as a result of the revised instructions. The largest difference is .03; most show no difference at all.

Table 2-1

Means and Standard Deviations of Instruction Times (in seconds) by Type of Instructions

	Instructions					
	Old			New		
Test	N	Mean	SD	N	Mean	SD
Assembling Objects	218	226.0	67.2	219	223.4	64.4
Orientation	218	284.7	78.4	219	282.2	70.3
Spatial Reasoning	217	108.6	42.0	219	102.6	35.4
One-Hand Tracking	218	182.3	44.7	219	185.7	42.7
Two-Hand Tracking	218	109.8	23.8	219	114.0	28.3
Target Identification	218	112.0	30.8	219	110.0	31.2

Table 2-2

Means and Standard Deviations of Test-Taking Times (in seconds) by Type of Instructions

	Instructions					
	Old			New		
Test	N	Mean	SD	N	Mean	SD
Assembling Objects	218	602.7	140.4	219	609.2	136.1
Orientation	218	317.5	101.9	219	319.8	101.5
Spatial Reasoning	217	462.9	162.7	219	475.2	162.8
One-Hand Tracking	217	225.3	3.3	219	225.9	22.6
Two-Hand Tracking	218	225.1	0.5	219	225.3	3.5
Target Identification	218	162.5	37.7	219	158.4	43.2

Table 2-3

Means and Standard Deviations of Spatial and Psychomotor/ Perceptual Speed Test Scores by Type of Instructions

	Instructions					
	Old			New		
Test	N	Mean	SD	N	Mean	SD
Assembling Objects	218	20.00	6.49	219	21.01	6.17
Orientation	218	11.53	5.88	219	11.66	5.76
Spatial Reasoning	215	18.19	6.22	213	18.84	6.23
One-Hand Tracking*	217	2.90	.42	217	2.91	.47
Two-Hand Tracking*	217	3.69	.49	219	3.67	.50
Target Identification						
Percent Correct	218	.93	.08	218	.93	.09
Decision Time	218	1.84	.61	218	1.78	.61
Movement Time	218	.36	.16	218	.36	.14

*Mean Log (Distance+1)

Table 2-4

Coefficients Alpha of Spatial Test Scores and Split-Half Reliabilities of Psychomotor/Perceptual Speed Test Scores by Type of Instructions

Test	Instructions	
	Old	New
Assembling Objects	.86	.85
Orientation	.87	.87
Spatial Reasoning	.87	.87
One-Hand Tracking	.95*	.95*
Two-Hand Tracking	.96*	.96*
Target Identification		
Percent Correct	.77*	.80*
Decision Time	.97*	.97*
Movement Time	.97*	.96*

*corrected using Spearman-Brown formula

Clarity Ratings. As indicated earlier, examinees rated the understandability of each test's instructions as being "very hard," "somewhat hard," or "not hard at all." These rating options were scored 1, 2, and 3, respectively. Table 2-5 shows means and standard deviations of the ratings provided by examinees who received the old or new instructions. With the exception of Spatial Reasoning, for which the old instructions were rated significantly less understandable ($t(434)=2.05$, $p<.05$), t-tests failed to reveal differences between the old and new instructions. Even for Spatial Reasoning, the old and new instructions are both rated as between "somewhat hard" and "not hard at all." The absolute difference, though statistically significant, is very small (2.37 vs. 2.51) on the 3-point scale.

Correlations with AFQT. The Armed Forces Qualification Test (AFQT) scores were available for the 207 examinees in the sample who were regular Army recruits. (They were not available for the Army Reserve or National Guard examinees.) Among the Army examinees, 103 were given the old instructions, and 104 received the new instructions. Table 2-6 reports the correlation between AFQT and instruction reading times for the spatial and psychomotor/perceptual speed tests. Fisher's z-test for comparing independent correlations (Hays, 1963) indicated that these correlations were not significantly different between the two sets of examinees.

Table 2-5

Means and Standard Deviations of Perceived Clarity Ratings by Type of Instructions

	Instructions					
	Old			New		
Test	N	Mean	SD	N	Mean	SD
Assembling Objects	218	2.68	0.57	219	2.68	0.58
Orientation	218	2.34	0.71	219	2.40	0.69
Spatial Reasoning	217	2.37	0.77	219	2.51	0.69
One-Hand Tracking	218	2.81	0.50	219	2.76	0.54
Two-Hand Tracking	218	2.72	0.63	219	2.75	0.57
Target Identification	218	2.89	0.41	219	2.86	0.46

Table 2-6

Correlations of AFQT with Instruction Time by Type of Instructions

	Instructions	
Test	Old	New
Assembling Objects	-.18	-.28
Orientation	-.30	-.32
Spatial Reasoning	-.26	-.13
One-Hand Tracking	-.25	-.05
Two-Hand Tracking	.11	.03
Target Identification	-.25	-.12

Table 2-7 shows the correlations between AFQT and test scores on the spatial and psychomotor/perceptual speed tests. Again, no significant differences were found between the correlations based on examinees who received the old and new instructions, respectively.

Table 2-7

Correlations of AFQT with Test Scores by Type of Instructions

Test	Instructions	
	Old	New
Assembling Objects	.29	.39
Orientation	.31	.40
Spatial Reasoning	.41	.39
One-Hand Tracking	-.25	-.31
Two-Hand Tracking	-.34	-.37
Target Identification		
Percent Correct	.04	.00
Decision Time	-.32	-.24
Movement Time	.06	.13

Summary. There were no differences in mean scores on the spatial and psychomotor/perceptual tests as a function of old or revised instructions. Furthermore, there were no differences with respect to either the time to read the instructions or the time to take the tests, and only one difference with respect to examinees' ratings of instructional clarity. Likewise, the reliabilities of the test scores, and the correlations between the test scores and AFQT, appear to have been unaffected by the revised instructions examined in this investigation.

Given the previous revisions that instructions for these tests had undergone, it is perhaps not surprising that almost no differences were found. There apparently was little room for additional improvement.

Results: Order of ABLE Items

Time to Complete. There was not a statistically significant difference between the two item orders for time to complete the ABLE ($p>.05$). The mean completion time for the original order was 1626.6 seconds (s.d.=422.5), and the mean time for reversed order was 1592.7 seconds (s.d.=402.6).

ABLE Scale Scores. Table 2-8 shows the means and standard deviations of the 14 ABLE scale scores (excluding the Non-Random Response scale) for examinees who received the original or reversed order of items. Student t-tests indicated significant mean differences due to item order on three of the 14 scales: Nondelinquency ($t(433)=2.10$, $p<.05$); Traditional Values

($t(433)=2.05$, $p<.05$); and Work Orientation ($t(433)=2.66$, $p<.01$). For these scales, the means were slightly higher when the ABLE was administered with the original item order. The effect sizes, computed using pooled variance, were .22 (Nondelinquency), .21 (Traditional Values), and .26 (Work Orientation). Additionally, the variance associated with the Fake Good (Social Desirability) scale was statistically significantly greater when the ABLE was administered with the reversed items ($F(216, 217)=1.33$, $p<.05$).

Table 2-8

Means and Standard Deviations of ABLE Scale Scores by Item Order

	Item Order					
	Original			Reversed		
ABLE Scale	N	Mean	SD	N	Mean	SD
Emotional Stability	216	40.53	6.38	219	40.47	6.02
Self-Esteem	216	29.82	3.98	219	29.33	4.22
Cooperativeness	216	44.64	5.15	219	43.73	5.45
Conscientiousness	216	37.20	4.42	219	36.39	4.29
Nondelinquency	216	47.61	5.59	219	46.30	6.16
Traditional Values	216	29.08	3.00	219	28.39	3.45
Work Orientation	216	46.45	6.56	219	44.79	6.27
Internal Control	216	42.98	4.09	219	42.48	4.32
Energy Level	216	51.29	6.60	219	50.53	6.43
Dominance	216	27.97	4.79	219	27.33	4.60
Physical Condition	216	14.28	2.76	219	13.98	3.15
Fake Good	216	16.90	3.54	219	16.44	3.08
Self-Knowledge	216	25.97	3.35	219	26.42	3.31
Fake Bad	216	1.19	1.90	219	1.00	1.75

Scale Reliabilities. Table 2-9 reports the reliabilities (coefficients alpha) for the ABLE scale scores for examinees who received the items in original and reversed orders. Generally, these reliabilities do not appear to differ by much. On average, the coefficients are about two points higher for the reversed item order (.76 versus .78). Two of the largest differences, for Traditional Values (.62 versus .71, for original and reversed orders, respectively) and Fake Good (.69 versus .63, for original and reversed orders, respectively) show differences in opposite directions.

Summary and Discussion. There was no difference in time to complete the ABLE across the two orders. The statistically significant, mean score differences were small and occurred for only three of fourteen scales. Nevertheless, we examined the changes to item order for each scale to see if these might explain the pattern of mean score differences. For each scale, we counted the number of items appearing on each half of the inventory. For seven of the scales, item appearances were balanced, with no more than 60% of the items appearing in either the front or back half of the inventory. One would probably not expect mean score differences for these scales across the original and revised item orders, since approximately equal numbers of items appear in front and back halves of the inventory regardless of order. There was a significant mean differences for one of these seven, Work Orientation, in favor of the original order.

Table 2-9

Coefficients Alpha of ABLE Scale Scores by Item Order

	Item Order	
Test	Original	Reversed
Emotional Stability	.79	.79
Self-Esteem	.79	.83
Cooperativeness	.77	.82
Conscientiousness	.77	.75
Nondelinquency	.77	.82
Traditional Values	.62	.71
Work Orientation	.89	.87
Internal Control	.74	.76
Energy Level	.87	.85
Dominance	.85	.84
Physical Condition	.80	.85
Fake Good	.69	.63
Self-Knowledge	.61	.67
Fake Bad	.73	.73

For the other seven scales, five had more than 60% of their items in the first half of the inventory ("front-loaded") when items were in the original order and more than 60% in the second half ("back-loaded") when items were in the reversed order. One of these scales, Traditional Values, showed a significant mean score difference, again in favor of the original order.

Two of the scales had more than 60% of their items in the second half of the originally ordered inventory and, thus, 60% in the first half of the reversed inventory. One of these scales, Nondelinquency, showed a significant mean score difference, once again in favor of the original order.

We conclude from this analysis that there is no apparent systematic relationship between item order and score differences on the ABLE scales. All three of the statistically significant mean score differences were in favor of the original item order, yet these differences occurred for one scale that had items balanced across inventory halves, and thus changed very little across order, one scale that changed from a "front-loaded" to a "back-loaded" scale across item orders, and one scale that changed from a "back-loaded" to a "front-loaded" scale across orders.

Finally, we reiterate the finding that the statistically significant mean differences were relatively small (no more than .26 effect size) and note that only two examinees failed the Non-Random Response screen -- one for each item order. Item order for the ABLE, as investigated here, does not appear to be a major concern.

CHAPTER 3

ASSESSING THE EQUIVALENCE OF COMPUTERIZED AND PAPER-AND-PENCIL TESTS OF TEMPERAMENT AND SPATIAL ABILITY

In Project A, measures of spatial ability and temperament were administered in a paper-and-pencil (P&P) format. However, as described in Chapters 1 and 2, some of these measures have since been adapted for computer administration. In particular, three of the spatial ability tests developed for Project A (Assembling Objects, Orientation, and Spatial Reasoning) are being administered on the Hewlett-Packard (HP) Integrated PC as part of the ECAT test battery. Likewise, the Army has administered a computerized version of the Assessment of Background and Life Experiences (ABLE) on the Seequa system (Walker, 1989), and this version has been transported for HP administration.

In the future, these spatial ability and temperament measures may continue to be administered on the computer. Therefore, it is important that scores on these measures be comparable across the two modes of administration. This is especially true if the computer-administered (CA) measures are ever to be fielded simultaneously with the P&P measures.

As described in Chapter 2, computerized versions of the three spatial ability tests and the ABLE were administered for this project at Fort Benning using the ECAT computers. In this chapter, we compare these data to data collected from the Project A longitudinal validation (LV) sample using the P&P versions of the instruments.

Samples

Two samples were used in this investigation. The CA sample (described in Chapter 2) is a subset of the 437 examinees from whom data were collected at the Fort Benning, Georgia, Reception Battalion. Recall that these examinees received the HP computer-administered versions of the three spatial tests and the ABLE. To maximize comparability with the LV sample (described below), only the 205 examinees who were regular Army recruits in the 11X Infantry Military Occupational Specialty (as opposed to Army Reserve, National Guard, or other recruits) were included in the present set of analyses.

The P&P sample is a subset of the nearly 50,000 Army recruits in the Project A LV sample from whom predictor data had been collected in 1986 and 1987. A random subsample of 823 Infantry recruits who received the P&P versions of the spatial tests and the ABLE was selected for this study.

Average Armed Forces Qualification Test (AFQT) scores were examined to determine whether the two samples were comparable in ability. The average AFQT scores for the examinees in the P&P and CA samples were 59.25 (s.d.=20.06) and 59.02 (s.d.= 17.74), respectively (t=-0.19, n.s.).

Measures

Paper-and-pencil measures. Project A examinees received P&P versions of a spatial test battery and the ABLE as part of a larger test battery. The Project A spatial test battery consisted of six separately-timed tests that required approximately one hour to administer. These tests were Assembling Objects, Map, Orientation, Reasoning, Maze, and Object Rotation. The spatial tests were administered as separate booklets and always in the same order. Test instructions were read by proctors, who also monitored the test-taking time limits. The time limits (in minutes) for the six tests were 16, 12, 10, 12, 5.5, and 7.5, respectively. The examinees answered multiple-choice questions by writing directly on the booklets. Of the six spatial tests, only data from the Assembling Objects, Orientation, and Reasoning tests are included in this investigation.

Computer-administered measures. Examinees at the Fort Benning, Georgia, Reception Battalion received the HP computer-administered versions of the three spatial tests included in the ECAT test battery and an HP version of the ABLE developed especially for this project. The spatial tests were self-paced and were always administered in the following order: (1) Assembling Objects, (2) Orientation, and (3) Spatial Reasoning. In contrast to the P&P version of Assembling Objects (which had a 16-minute time limit), the CA version had a 12-minute time limit. The time limits for Orientation and Spatial Reasoning were 10 and 12 minutes, respectively (the same as for the P&P versions). As described in Chapter 2, examinees at Fort Benning received either of two sets of instructions with these measures. The instructions in one set were those adapted from the Project A instructions for ECAT testing. These instructions were intended to be the same as those used in Project A (however, see Chapter 2 for a summary of differences). Instructions in the second set were revised (for improved readability) versions of the first set. Analyses reported in Chapter 2 indicated that instruction set had no impact on test score characteristics; therefore, this distinction is ignored in the analyses reported in the present chapter.

The ABLE was self-paced and administered with no time limit. Examinees received the ABLE items in one of two orders: the original order used in Project A, or reversed. Results of analyses reported in Chapter 2 indicated that there were a few mean scale score differences associated with item order; however,

these differences were small and nonsystematic. Therefore, this distinction is also ignored in the analyses reported in this chapter.

The items on the CA versions of the Spatial Reasoning and Orientation tests were administered in the same order as those on the P&P versions. However, as documented in Table 3-1, the items and item order on the Assembling Objects test differed somewhat between the CA and P&P versions. The CA version of the Assembling Objects test was the 32-item test used during the Project A concurrent validation data collection. The Project A P&P LV version had 36 items. Column 3 of Table 3-1 delineates the discrepancies between the two tests. Analyses of the Assembling Objects test in this chapter are performed using the 28 items common to both test versions. Additionally, as described in Chapter 2, five items from the ABLE were treated as missing due to text errors in the CA version of the instrument. However, scale scores were adjusted using missing data scoring rules developed in Project A, so that the maximum possible score for each scale was not affected.

The difference in time limits of the CA and P&P versions of the Assembling Objects test also complicates the interpretation of any observed score differences between them. In effect, there was about a 15% reduction in time allowed for the CA version of the Assembling Objects test (about 26.6 seconds per item for the 36-item P&P version versus about 22.5 seconds per item for the 32-item, CA version).

Analyses

Spatial tests. Means and standard deviations were computed for the CA and P&P versions of each of the three tests. Student t-tests (Hays, 1963) were computed to test for significant mean differences between the CA and P&P versions. Comparisons between the CA and P&P versions were also made with respect to the internal consistency reliabilities of the three tests, the intercorrelations among the tests, and the correlations of the tests with AFQT. Additionally, the item difficulty characteristics of the CA and P&P versions were compared.

ABLE. Proportions of examinees screened by the Non-Random Response scale were compared for the CA and P&P versions of the ABLE, as were means and standard deviations for each of the other 14 scales. A chi-square test was used to compare the CA and P&P proportions, and Student t-tests were used to compare the means of the CA and P&P scales. Comparisons between the CA and P&P versions were also made with respect to the internal consistency reliabilities of the scales, the intercorrelations among the scales, and the correlations of the scales with AFQT. Finally, factor analyses of the 11 substantive scale scores were computed for both the CA and P&P versions.

Table 3-1

Translation Between Computer-Administered (32 items) and Paper-and-Pencil (36 items) Assembling Objects Items

CA (32 Items)	P&P (36 Items)	Description
1	1	Match
2	2	Match
3	3	Changed stem and options
4	4	Changed stem and options
5	5	Match
6	6	Match
7	7	Match
8	8	Match
9	9	Match (except lettering on stem)
10	10	Match
	11	
	12	
	13	
11	14	Match
12	15	Match
13	16	Match (except lettering on stem)
14	17	Match
15	18	Match
16		
17	24	Changed stem and options
18	25	Match
19	26	Match
20	27	Match
21	19	Match
22	20	Match
23	21	Match
24	22	Match
25	23	Match
26	28	Match
27	29	Match
28	30	Match
29	31	Match
30	32	Match
31	33	Match
32	34	Match
	35	
	36	

Results: Spatial Tests

Means and Standard Deviations. Table 3-2 reports the means and standard deviations for total (number correct) scores for each of the three spatial tests. These results are reported separately for examinees who received the P&P and CA versions of the tests, respectively. The results indicate that examinees who received the P&P versions of the tests achieved significantly higher scores than examinees who received the CA versions on both the Assembling Objects ($t(991)=5.36$, $p<.001$) and Spatial Reasoning ($t(285.4)=2.86$, $p<.01$) tests. The effect sizes of these differences were approximately two-fifths of a standard deviation for Assembling Objects, and approximately one-fourth of a standard deviation for Spatial Reasoning. Additionally, the variance associated with the CA version of the Spatial Reasoning test was significantly greater than for the P&P version ($F(201, 791)=1.29$, $p<.05$).

Table 3-2

Means and Standard Deviations of Spatial Test Scores by Method of Administration

	Method of Administration					
	P&P			CA		
Test	N	Mean	SD	N	Mean	SD
Assembling Objects	788	20.10	5.71	205	17.71	5.56
Orientation	791	12.51	6.26	205	11.85	5.66
Spatial Reasoning	792	20.13	5.15	202	18.84	5.86

Reliabilities. Table 3-3 reports the internal consistency reliabilities (coefficients alpha) of the three spatial tests. As in Table 3-2, these results are reported separately for examinees who received the P&P and CA versions of the tests. The coefficients alpha are relatively high for both versions of each of the spatial tests, and there is little difference between the two versions of the tests.

Table 3-3

Coefficients Alpha of Spatial Test Scores by Method of Administration

Test	Method of Administration P&P	CA
Assembling Objects	.87	.83
Orientation	.89	.86
Spatial Reasoning	.83	.85

Test Score Intercorrelations and Correlations with AFQT. Table 3-4 reports the correlations among the three spatial tests and between each of the spatial tests and AFQT. The upper-right triangle shows the correlations for examinees who received the P&P versions of the tests, and the lower-left triangle shows the correlations for examinees who received the CA versions. Fisher z-tests between corresponding elements of the upper- and lower-triangles failed to indicate significant differences between the P&P and CA test versions.

Table 3-4

Spatial Test Score Intercorrelations and Correlations with AFQT by Method of Administration

	AO	OR	SR	AFQT
AO	-	.44	.51	.35
OR	.44	-	.45	.40
SR	.53	.47	-	.51
AFQT	.34	.36	.42	-

Note: The correlations above the diagonal are for the P&P test scores (n=786-789); the correlations below the diagonal are for the CA test scores (n=203-204).

Item Difficulties. The preceding analyses indicate that the internal consistency reliabilities of the P&P and CA versions of the three spatial tests are very similar to one another, as are their intercorrelations and correlations with AFQT. Still, comparisons of the total number correct scores (Table 3-2) did indicate that examinees who received the P&P versions of the Assembling Objects and Spatial Reasoning tests scored higher (on average) than examinees who received the corresponding CA versions. For the Assembling Objects test, this could perhaps be explained, at least in part, by the greater time limit for the P&P version. However, the CA version of the Spatial Reasoning test had the same time limit as the P&P version, and it still showed a significantly lower mean score. To better understand these mean score differences, we examined the item difficulties for the tests.

The proportion of examinees providing the correct response to each item is reported in Tables 3-5, 3-6, and 3-7 for each of the three spatial tests, respectively. Asterisks in the P&P column indicate that the item was answered correctly by a significantly greater proportion of examinees who received that version of the test than by examinees who received the CA version. Conversely, asterisks in the CA column indicate that the item was answered correctly by a greater proportion of examinees who received the test via the computer. Generally, the results are very similar to the test score comparisons reported above. That is, the greatest number of differences were found for items on the Assembling Objects test, followed by items on the Spatial Reasoning Test. Altogether, 17 of the 28 Assembling Objects items and 12 out of the 30 Spatial Reasoning items were significantly more difficult when computer-administered. In contrast, only four out of 24 Orientation items were significantly more difficult on the computer.

If the amount of time allowed for completing the tests was the primary reason for the Assembling Objects results, then we would expect to see: (1) equally difficult items in the first half of the CA and P&P versions of the Assembling Objects test, but more difficult items in the second half of the CA version of the test relative to the P&P version; and (2) no differences in the item difficulties between the CA and P&P versions of the Orientation and Spatial Reasoning tests in either half. Table 3-8 summarizes the results in Table 3-5 through 3-7 to help examine these hypotheses.

Table 3-5

Proportion Correct for Assembling Objects Items by Method of Administration

	Method of Administration	
Item Number	CA	P&P
01	80	82
02	69	80*
05	75	79
06	71	79*
07	82	82
08	65	81*
09	73	81*
10	55	68*
11	66	73
12	50	48
13	49	68*
14	59	62
15	61*	50
18	87*	81
19	77	78
20	68	77*
21	65	89*
22	73	81*
23	71	83*
24	70	79*
25	66	76*
26	51	64*
27	32	68*
28	67	74*
29	56	61
30	43	53*
31	53	59
32	39	53*

Note: Asterisks in the column headed CA indicate that the proportion correct for an item was significantly greater ($p<.05$) when administered on the computer. Asterisks in the column headed P&P indicate that the proportion correct was significantly greater when the item was administered by paper-and-pencil. Decimals have been omitted from proportions.

Table 3-6

Proportion Correct for Orientation Items by Method of Administration

	Method of Administration	
Item Number	CA	P&P
01	42	47
02	37	48*
03	47	57*
04	52	49
05	38	45
06	61	58
07	61	67
08	22	19
09	43	53*
10	57	58
11	46	54*
12	37	41
13	55	60
14	57	57
15	65	70
16	50	54
17	24	26
18	80	79
19	42	50
20	44	42
21	57	57
22	59	58
23	57	58
24	53	46

Note: Asterisks in the column headed P&P indicate that the proportion correct was significantly greater ($p<.05$) when the item was administered by paper-and-pencil. Decimals have been omitted.

Table 3-7

Proportion Correct for Spatial Reasoning Items by Method of Administration

	Method of Administration	
Item Number	CA	P&P
01	82	91*
02	85	92*
03	79	86*
04	87	93*
05	85	89
06	82	89*
07	71	73
08	61	66
09	75	83*
10	54	48
11	45	61*
12	47	43
13	30	39*
14	42	43
15	36	40
16	74	88*
17	74	78
18	78	87*
19	77	83
20	80	89*
21	72	80*
22	74	75
23	74	78
24	44	58
25	43	50
26	64	63
27	61	55
28	53	47
29	22	19
30	29	30

Note: Asterisks in the column headed P&P indicate that the proportion correct was significantly greater ($p<.05$) when the item was administered by paper-and-pencil. Decimals have been omitted.

Table 3-8

Mean Item Difficulty Indexes for First and Second Halves of Computer-Administered and Paper-and-Pencil Versions of Three Spatial Tests by Method of Administration

	Test Half					
	First Half			Second Half		
	CA	P&P	CA-P&P	CA	P&P	CA-P&P
AO	67.29	72.43	-5.14	59.36	71.07	-11.71
OR	45.25	49.67	-4.42	53.58	54.75	-1.17
SR	64.07	69.07	-5.00	61.27	65.33	-4.06

Not surprisingly, the results in Table 3-8 indicate that the CA versions of the Assembling Objects and Spatial Reasoning tests are generally more difficult than the P&P versions. However, so is the first half of the Orientation test (recall that the difference between Orientation total scores was not statistically significant). Our hypothesis appears to be confirmed for Assembling Objects in that the second half of the CA version is about 11 points more difficult than the P&P version -- which is about double the difference shown for the first halves (about 5 points). Such results were not found for the Spatial Reasoning test. The difference between the second halves of the two versions (a 4-point difference) is about the same as between the first halves (a 5-point difference). We conclude from these findings that the differential amounts of time allowed for the P&P and CA test versions probably accounts for some, though not all, of the score differences on the Assembling Objects test.

An examination of the items identified as more difficult on the CA version of the Assembling Objects and Spatial Reasoning tests indicates that most of these items use drawings of complex shapes, such as drawings of three-dimensional shapes or drawings with many small pieces. These items appear to be more difficult when administered by computer than when administered using paper and pencil. Tests of spatial ability are particularly dependent on how well the items are graphically represented, either as displayed on a computer screen or drawn on paper. For example, straight lines drawn at certain angles on paper may appear jagged on the computer screen.

Results: ABLE

Proportions of Examinees Screened by Non-Random Response Scale. As described in Chapter 2, one of the ABLE scales -- the Non-Random Response scale -- was developed to screen out examinees who either do not (or cannot) read the questions, and respond randomly. Table 3-9 reports the number of examinees identified as randomly responding on the ABLE for both the CA and P&P versions. A comparison of these numbers indicates that a significantly larger proportion of examinees who received the P&P ABLE were identified as random responders than were examinees who received the CA ABLE (chi-square(1)=6.07, $p<.05$). All examinees identified as random responders were removed from the remaining ABLE analyses reported in this chapter. (Likewise, no further analyses were conducted on that scale.)

Table 3-9

Examinees Identified as Random Responders Using the ABLE Non-Random Response Scale by Method of Administration

Method of Administration	Total Number of Examinees	Examinees Identified as Random Responders
CA	205	2
P&P	823	39

Means and Standard Deviations. Table 3-10 reports the means and standard deviations for each of the ABLE scales (other than the Non-Random Response scale). Student t-tests indicated that examinees who received the ABLE on the computer scored significantly higher than examinees who received the P&P ABLE on only one scale (Traditional Values; $t(282)=2.79$, $p<.01$). The size of this difference was approximately one-fifth of a standard deviation. On the other hand, it should be noted that the variances associated with the CA ABLE were significantly greater ($p<.05$) than those associated with the P&P ABLE for seven of the 11 substantive scales (Emotional Stability, Self-Esteem, Cooperativeness, Traditional Values, Work Orientation, Energy Level, and Dominance) and one of the three remaining validity scales (Fake Bad).

Table 3-10

Means and Standard Deviations of ABLE Scale Scores by Method of Administration

	Method of Administration					
	Computerized			Paper-and-Pencil		
Scale	N	Mean	SD	N	Mean	SD
Emotional Stability	203	41.21	6.66	782	41.24	5.52
Self-Esteem	203	29.73	4.52	782	29.50	3.92
Cooperativeness	203	43.96	5.57	782	44.67	4.73
Conscientiousness	203	36.89	4.30	782	37.10	3.99
Nondelinquency	203	47.20	5.91	782	47.31	5.56
Traditional Values	203	28.61	3.35	780	29.35	2.82
Work Orientation	203	45.57	6.69	782	45.97	5.98
Internal Control	203	43.03	4.45	782	42.51	4.12
Energy Level	203	51.43	7.02	782	51.94	5.86
Dominance	203	27.64	5.13	782	28.29	4.50
Physical Condition	203	14.15	3.10	782	14.17	2.85
Fake Good	203	16.66	3.25	782	16.81	3.31
Self-Knowledge	203	26.34	3.40	782	26.36	3.18
Fake Bad	203	1.07	1.89	782	1.09	1.58

Reliabilities. Table 3-11 reports the internal consistency reliabilities (coefficients alpha) of the fourteen ABLE scales (not including the Non-Random Response scale). The internal consistencies of the CA ABLE scales are uniformly higher than the internal consistencies of the P&P scales. The average internal consistency reliability of the 11 substantive scales on the ABLE was .82 for the CA version and .76 for the P&P version.

Table 3-11

Coefficients Alpha of ABLE Scale Scores by Method of Administration

Scale	Method of Administration	
	CA	P&P
Emotional Stability	.82	.74
Self-Esteem	.85	.78
Cooperativeness	.82	.75
Conscientiousness	.75	.72
Nondelinquency	.80	.75
Traditional Values	.68	.59
Work Orientation	.89	.85
Internal Control	.79	.73
Energy Level	.89	.82
Dominance	.87	.82
Physical Condition	.86	.79
Fake Good	.65	.63
Self-Knowledge	.64	.61
Fake Bad	.75	.63

Scale Score Intercorrelations and Correlations with AFQT. Table 3-12 reports the correlations among the ABLE scale scores and between each scale score and the AFQT. Results for the P&P ABLE are reported in the upper-right triangle, and results for the CA ABLE are reported in the lower-left triangle. Fisher z-tests indicate that 31 of the 55 correlations among the 11 ABLE substantive scales (those in the first 11 rows and columns of the correlation matrix) were significantly higher ($p<.05$) for examinees who received the CA as opposed to the P&P ABLE. On the other hand, only one of the 33 correlations between the 11 substantive scales and the three validity scales (Fake Good, Self-Knowledge, and Fake Bad) was moderated by method of administration. None of the correlations between the ABLE scales and AFQT was significantly different across the two samples.

Table 3-12

ABLE Scale Score Intercorrelations and Correlations with AFQT by Method of Administration

Scale	ES	SE	Coop	Cons	M	TV	WO	IC	EL	D	PC	PG	SK	PB	AFQT
Emotional Stability	-	.60	.53	.45	.35	.26	.55	.44	.60	.53	.34	.24	-.02	-.63	.15
Self-Esteem	.75	-	.46	.47	.33	.23	.64	.36	.70	.63	.44	.21	.16	-.45	.23
Cooperativeness	.58	.56	-	.43	.48	.38	.47	.40	.52	.37	.19	.29	.11	-.45	.08
Conscientiousness	.61	.63	.61	-	.53	.43	.60	.45	.59	.45	.27	.39	.18	-.39	.02
Nondelinquency	.49	.37	.59	.67	-	.53	.42	.35	.38	.21	.18	.42	.07	-.38	.08
Traditional Values	.47	.44	.54	.60	.68	-	.39	.34	.34	.23	.14	.30	.07	-.29	-.05
Work Orientation	.67	.72	.58	.77	.58	.52	-	.46	.77	.60	.43	.38	.21	-.41	.06
Internal Control	.56	.50	.54	.52	.49	.54	.57	-	.48	.33	.15	.12	.09	-.42	.14
Energy Level	.81	.76	.59	.71	.54	.54	.85	.58	-	.57	.51	.30	.12	-.54	.11
Dominance	.63	.73	.39	.53	.23	.33	.59	.39	.62	-	.34	.22	.19	-.36	.17
Physical Condition	.48	.56	.33	.39	.26	.17	.59	.28	.63	.43	-	.08	.18	-.26	.02
Fake Good	.35	.30	.40	.51	.49	.36	.42	.14	.36	.22	.20	-	.08	-.11	-.20
Self-Knowledge	.09	.20	.07	.15	.00	.07	.21	.17	.20	.20	.18	-.06	-	.02	.04
Fake Bad	-.70	-.55	-.47	-.44	-.44	-.42	-.50	-.52	-.62	-.44	-.41	-.13	-.02	-	-.17
AFQT	.18	.17	.11	.05	.10	.04	.10	.19	.16	.15	.10	-.05	.08	-.11	-

Note: The correlations above the diagonal are for the paper-and-pencil ABLE (n=780-782); the correlations below the diagonal are for the computerized ABLE (n=203). Underlined correlations below the diagonal are significantly higher than the corresponding correlations above the diagonal at the $p<.05$ level.

Factor Analyses of Scale Scores. Factor analyses of the 11 substantive scale scores for both the P&P and CA ABLE were also conducted. Principal axis factor solutions were determined using squared multiple correlations as the initial communality estimates. The initial unrotated results indicated that the amount of common variance associated with the CA version of the ABLE was greater than that associated with the P&P version of the instrument (7.07 vs. 5.54). A comparison of the unrotated eigenvalues from the two solutions with the eigenvalues resulting from a parallel analysis suggests that no more than three factors should be extracted for the CA data, but as many as five factors could be extracted from the P&P data. (Parallel analysis is a technique which does not identify the exact number of factors to be extracted, but does identify the point at which further factor extraction will likely result in the interpretation of random data [Humphreys & Montanelli, 1975; Montanelli & Humphreys, 1976].) Prior research on the ABLE indicates that anywhere from two to three common factors can be identified from the ABLE scales. For example, results of principal factor analyses of ABLE scale score data collected in the Project A field test (Peterson, 1987) suggested two factors, labeled "Personal Impact" and "Dependability." (Note that those factor analyses did not include the Physical Condition scale.) Principal factor analyses of ABLE scale score data from the Project A longitudinal validation sample suggested that three factors could be extracted, but that only two factors seemed meaningful to interpret (Peterson, Russell, Hallam, Hough, Owens-Kurtz, Gialluca, & Kerwin, in press). In terms of the defining variables, the two interpretable factors were virtually identical to the Personal Impact and Dependability factors found in the field test.

Based on these considerations, we decided to retain and examine the two- and three-factor solutions. Results of varimax-rotated two-factor solutions for the CA and P&P versions of the ABLE are reported in Table 3-13. These results appear to be very similar across the two samples. For example, the six scales which load higher on the first factor are the same across the two solutions, as are the five scales which load higher on the second factor. These two factors appear to be tapping the same Personal Impact and Dependability factors identified in the Project A field test and longitudinal validation samples. Note that the factor loadings are generally greater in the factor solution for the CA ABLE scales than for P&P ABLE. This finding is reflected in the greater amount of common variance accounted for by the two factors in the CA data (7.14) than in the P&P data (5.80).

Table 3-13

Results of Principal Factor Analyses with Varimax Rotation by Method of Administration: Two-Factor Solutions

CA Scales (N=203)	Factor 1	Factor 2	h^2
Emotional Stability	.71	.46	.72
Self-Esteem	.81	.33	.77
Work Orientation	.69	.55	.78
Energy Level	.77	.51	.85
Dominance	.73	.20	.57
Physical Condition	.65	.13	.44
Cooperativeness	.38	.63	.54
Conscientious	.49	.68	.70
Nondelinquency	.16	.81	.68
Traditional Values	.19	.76	.61
Internal Control	.37	.57	.46
Eigenvalues	3.78	3.36	7.14

P&P Scales (N=780)	Factor 1	Factor 2	h^2
Emotional Stability	.68	.34	.58
Self-Esteem	.77	.25	.66
Work Orientation	.70	.48	.72
Energy Level	.80	.39	.79
Dominance	.67	.21	.49
Physical Condition	.52	.08	.28
Cooperativeness	.39	.54	.44
Conscientious	.43	.62	.57
Nondelinquency	.16	.71	.53
Traditional Values	.11	.64	.42
Internal Control	.34	.47	.34
Eigenvalues	3.39	2.41	5.80

Results of varimax-rotated three-factor solutions are reported in Table 3-14. Note that the five scales which loaded higher on the Dependability factor (i.e., Factor 2) in both of the two-factor solutions also load highest on the first factor of the CA three-factor solution and on the second factor of the P&P three-factor solution. On the other hand, the six scales which previously loaded higher on the Personal Impact factor (i.e., Factor 1) of the two-factor solutions all load highest on the first factor of the three-factor solution for the paper-and-pencil data, but are split between two factors (the second and third) of the three-factor solution for the CA data. Examination of the pattern with which the six scales load on these two factors suggests that the Personal Impact factor has been divided into two subfactors. These subfactors might be labeled "Energy" and "Self-Confidence." Finally, as was the case for the two-factor results, the amount of common variance accounted for by the three factors was greater for the CA data (7.40) than for the P&P data (6.04).

Discussion

In this section we summarize the findings from the analyses described above. We begin with a summary of the spatial results, and finish with a summary of the ABLE results.

Summary of Spatial Test Results. The three spatial tests are very similar across the CA and P&P versions with regard to internal consistency reliability, intercorrelations, and correlations with AFQT. Two of the tests, Assembling Objects and Spatial Reasoning, showed significantly lower scores for the CA versions. These differences could not be totally explained by differences in the time allowed to take the two versions of the test. Item difficulty analyses identified the items that appeared to be more difficult in the CA versions. These items appeared to be more complex with more small pieces. It appears that degradation in the clarity and resolution of these kinds of items in the computer presentation may account for much of the observed test score differences.

Summary of ABLE Scale Results. There are several differences between scale scores associated with the CA and P&P versions of the ABLE. First, the proportion of examinees identified as random responders was greater among those who received the P&P ABLE. Second, although mean differences were found on only one scale (i.e., Traditional Values), significant differences in variance were found on seven of the 11 substantive scales. The CA ABLE scores were always more variable than the P&P ABLE scores. Third, the internal consistency reliabilities were greater on the CA scales as opposed to the P&P scales.

Table 3-14

Results of Principal Factor Analyses with Varimax Rotation by Method of Administration: Three-Factor Solutions

CA Scales (N=203)	Factor 1	Factor 2	Factor 3	h^2
Emotional Stability	.44	.66	.33	.74
Self-Esteem	.31	.72	.41	.78
Work Orientation	.52	.41	.62	.82
Energy Level	.48	.54	.59	.87
Dominance	.18	.70	.30	.61
Physical Condition	.10	.35	.62	.52
Cooperativeness	.62	.34	.21	.54
Conscientiousness	.66	.35	.38	.70
Nondelinquency	.81	.07	.22	.71
Traditional Values	.75	.22	.07	.62
Internal Control	.56	.37	.16	.48
Eigenvalues	3.19	2.46	1.75	7.40

P&P Scales (N=780)	Factor 1	Factor 2	Factor 3	h^2
Emotional Stability	.54	.27	.52	.63
Self-Esteem	.70	.21	.35	.66
Work Orientation	.72	.48	.09	.76
Energy Level	.76	.36	.29	.79
Dominance	.63	.18	.26	.50
Physical Condition	.54	.09	.04	.30
Cooperativeness	.28	.48	.43	.49
Conscientiousness	.46	.62	.07	.60
Nondelinquency	.14	.69	.17	.52
Traditional Values	.11	.63	.08	.42
Internal Control	.28	.44	.27	.34
Eigenvalues	2.98	2.21	0.85	6.04

A fourth difference concerns the correlations among the scales. Generally, the level of correlations among the CA scales was greater than that among the P&P scales. In a related finding, results of factor analyses indicated that the amount of common variance associated with the CA scales was greater than that associated with the P&P scales. These results also indicated that the CA data supported a more finely differentiated factor structure than did the P&P data.

Several of these differences suggest that examinees who received the CA version of the ABLE may have taken the instrument more seriously and paid more attention when responding to the items. For example, such an hypothesis is consistent with the difference in proportions of examinees identified as random responders across the two versions of the ABLE. Likewise, if examinees were responding more carefully, we would expect increases in the reliability of the scale scores, as well as higher correlations among the scales (due to the increases in reliability). Furthermore, the more differentiated factor pattern of the computerized ABLE might also be expected if examinees were making a more concerted effort to respond accurately.

Despite all of these differences between the CA and P&P ABLE, it is important to note the lack of differences in mean scores on the scales. The Traditional Values scale was the only one on which the average scores were significantly different (i.e., CA greater). We know from prior research (Peterson, 1987) that scores on the ABLE can be significantly raised or lowered if examinees are instructed to fake good or bad, respectively. Therefore, the findings reported here suggest that examinees who received the CA ABLE were not any more motivated to "fake" (e.g., respond in a socially desirable fashion) than were examinees who took the P&P ABLE.

CHAPTER 4

PRACTICE EFFECTS ON COMPUTERIZED TESTS OF PERCEPTUAL SPEED AND PSYCHOMOTOR ABILITY

Prior investigations using psychomotor and perceptual speed tests suggest that performance is likely to improve with practice. For example, Bilodeau (1952, cited in Jones, 1969) found that performance on the second trial on the Two-hand Coordination Test improved approximately three-quarters of a standard deviation over the first trial, and performance on the eighth trial improved almost two-and-a-half standard deviations. More recently, Ackerman (1988) reported improvements in performance of approximately one standard deviation between the first and second trials of a 60-item, nine-choice reaction time task.

The Project A psychomotor and perceptual speed tests were validated on examinees who had no prior experience with them. However, if these tests were to become operational, it is conceivable that future examinees would attempt to improve their performance by practicing beforehand. If practice on these tests is effective and were to be available (for instance, through a test preparation company), some examinees (those who practiced) would have an unfair advantage. Moreover, it is possible that the psychometric properties of the scores from these tests might change with practice. The purpose of the investigation reported in this chapter was to assess the effects of practice on scores from the computerized tests developed in Project A and included in the ECAT battery.

Due to resource limitations, only one of the two tracking tests, Two-Hand Tracking, was included in this research, along with Target Identification. Two-Hand Tracking was chosen for two reasons. First, prior research has demonstrated it to be the more difficult of the two tracking tests (e.g., Peterson, 1987). We believed that this might allow greater opportunity for practice effects to occur. Second, research reported by Busciglio, Silva, and Walker (1990) indicated that the validity of Two-Hand Tracking for predicting gunnery performance was greater than that of One-Hand Tracking.

Sample

Data were collected at the Fort Knox Reception Battalion from 230 male examinees with no previous military experience. Each examinee was assigned to one of two practice conditions: Two-Hand Tracking or Target Identification. A total of 114 examinees were assigned to the Two-Hand Tracking condition. Armed Services Vocational Aptitude Battery (ASVAB) scores were available for 112 of these examinees. The mean Armed Forces

Qualification Test (AFQT) score was 58.02, and the standard deviation was 17.05. The remaining 116 examinees were assigned to the Target Identification condition. ASVAB scores were available for 111 of these examinees. The mean AFQT score was 59.95, and the standard deviation was 17.59.

Measures

Depending on the condition to which they were assigned, examinees received either the Two-Hand Tracking test or the Target Identification test. Each test was administered five times, with a one-minute rest period between each test administration (or trial). For each test condition, items were presented in the same order during the first and last trial, but were scrambled during the second, third, and fourth trials. Also, test instructions were administered prior to the first trial only. Detailed scoring methods for the Two-Hand Tracking and the Target Identification tests are described in Chapter 2.

Analyses

Means, standard deviations and reliabilities were computed for each trial for all test scores. Repeated measures analysis of variance and t-tests for dependent samples were used to compare the means of adjacent test trials. To determine if practice effects were moderated by cognitive ability, soldiers were divided into three categories by AFQT score. Means and standard deviations of the Two-Hand Tracking and Target Identification test scores were computed for soldiers in each of these categories. Analyses of variance and t-tests were again computed where appropriate.

Results

Two-Hand Tracking. Means, standard deviations, and split-half (odd-even) reliabilities of mean log distance scores for examinees with scores for each of the five test trials are reported in Table 4-1. The table indicates that the mean of the mean log distance scores decreased across trials, the standard deviation increased, and the split-half reliability remained high and relatively unchanged (r_{xx} = .97-.98).

Repeated measures analysis of variance (ANOVA) results indicated that the differences in mean test scores across trials were statistically significant ($F(4,444)=177.64$, $p<.001$). Results of paired-comparison t-tests comparing means of adjacent test trials are reported in Table 4-2. These results demonstrate that mean performance improved significantly through the fourth trial, but that the difference between the fourth and fifth trial was not significant.

Table 4-1

Descriptive Statistics for Mean Log Distance Score by Trial (N=112)

Trial	Mean	SD	Odd-Even r_{xx}*
1	3.57	0.43	0.97
2	3.34	0.45	0.97
3	3.17	0.46	0.98
4	3.10	0.50	0.98
5	3.10	0.51	0.98

*corrected for length using Spearman-Brown

Table 4-2

Paired-Comparison t-Tests for Mean Change in Mean Log Distance Score Between Adjacent Trials

Trials	Mean Change	SD	t	p-value
1 and 2	0.23	0.02	12.67	0.0001
2 and 3	0.17	0.02	10.24	0.0001
3 and 4	0.07	0.02	4.49	0.0001
4 and 5	0.00	0.02	0.18	0.8557

To determine if the practice effects reported above were moderated by cognitive ability, the sample was divided into three subgroups according to their AFQT category. The range of AFQT percentile scores associated with each category is shown in Table 4-3. Because only two examinees in the sample belonged to Category 1, they were combined into one group with the 40 examinees belonging to Category 2. Of the remaining examinees for whom ASVAB scores were available, 35 belonged to Category 3A, and 35 belonged to Category 3B. Means and standard deviations of mean log distance scores computed separately by AFQT category and test trial are reported in Table 4-4.

Table 4-3

AFQT Categories

AFQT Category	Range of AFQT Percentile Scores
1	93 - 99
2	65 - 92
3A	50 - 64
3B	31 - 49

Table 4-4

Descriptive Statistics for Mean Log Distance Score by AFQT Category and Trial

AFQT Category (n)	Trial	Mean	SD
1 & 2 (42)	1	3.42	0.39
	2	3.20	0.41
	3	3.03	0.41
	4	2.97	0.49
	5	2.96	0.48
3A (35)	1	3.54	0.40
	2	3.30	0.41
	3	3.13	0.42
	4	3.05	0.46
	5	3.03	0.43
3B (35)	1	3.79	0.41
	2	3.55	0.48
	3	3.39	0.49
	4	3.31	0.50
	5	3.33	0.55

The results in Table 4-4 indicate that the mean log distance scores differed systematically across the three AFQT category subgroups. Specifically, holding test trial constant, scores for examinees in Categories 1 and 2 (combined) were consistently lower than scores for examinees in Category 3A, which were

consistently lower than scores for examinees in Category 3B. The results in Table 4-4 also indicate that mean test scores decreased across test trials for examinees in each of the three groups, and that the pattern of change was approximately the same in each.

Between-subjects repeated measures ANOVA results support these observations. Both the between-subjects effects of AFQT category ($F(2,109)=7.10$, $p<.01$) and the within-subjects effect of test trial ($F(4,436)=174.48$, $p<.001$) were statistically significant. However, the interaction between AFQT category and trial was not significant ($F(8,436)=0.20$, n.s.). Results of paired-comparison t-tests, reported in Table 4-5, indicate that mean performance for all three subgroups improved significantly across each of the first four trials, but did not change significantly between the fourth and fifth trials. Displayed graphically in Figure 4-1, these results suggest that the "practice effect curves" for the three subgroups were parallel, but occurred at different levels of elevation.

Target Identification: Clipped Mean Decision Time. Means, standard deviations, and split-half (odd-even) reliabilities of clipped mean decision time scores for examinees with scores for each of the five trials on the Target Identification test are reported in Table 4-6. Similar to the Two-Hand Tracking results reported above, the results reported in Table 4-6 indicate that mean performance improved across trials and the split-half reliability remained high and relatively unchanged (r_{oe} = .97-.98). However, in contrast to the Two-Hand Tracking results, the standard deviation of the clipped mean decision time scores decreased (rather than increased) across trials.

Repeated measures ANOVA results indicated that the differences in mean decision time scores across trials were statistically significant ($F(4,452)=176.99$, $p<.001$). Results of paired-comparison t-tests comparing means of adjacent test trials are reported in Table 4-7. These results show that mean performance improved significantly across all five trials.

Table 4-5

Paired-Comparison t-Tests for Mean Change in Mean Log Distance Score Between Adjacent Trials by AFQT Category

AFQT Category	Trials	Mean Change	SD	t	p-value
1 & 2	1 and 2	0.22	0.03	7.48	0.0001
	2 and 3	0.17	0.03	6.36	0.0001
	3 and 4	0.07	0.02	2.86	0.0067
	4 and 5	0.01	0.02	0.35	0.7310
3A	1 and 2	0.24	0.02	10.08	0.0001
	2 and 3	0.17	0.03	5.97	0.0001
	3 and 4	0.08	0.02	3.27	0.0024
	and 5	0.00	0.03	0.72	0.4780
3B	1 and 2	0.23	0.04	5.79	0.0001
	2 and 3	0.16	0.03	5.27	0.0001
	3 and 4	0.08	0.04	2.06	0.0470
	4 and 5	0.00	0.03	-0.72	0.4767

Table 4-6

Descriptive Statistics for Clipped Mean Decision Time Score by Trial (N=114)

Trial	Mean	SD	Odd-Even r_{xx}*
1	1.71	0.63	0.98
2	1.33	0.53	0.98
3	1.20	0.44	0.97
4	1.08	0.40	0.98
5	0.99	0.32	0.98

*corrected for length using Spearman-Brown

Table 4-7

Paired-Comparison t-Tests for Mean Change in Clipped Mean Decision Time Score Between Adjacent Trials

Trials	Mean Change	SD	t	p-value
1 and 2	0.38	0.03	12.34	0.0001
2 and 3	0.13	0.02	6.90	0.0001
3 and 4	0.12	0.02	7.85	0.0001
4 and 5	0.09	0.02	5.10	0.0001

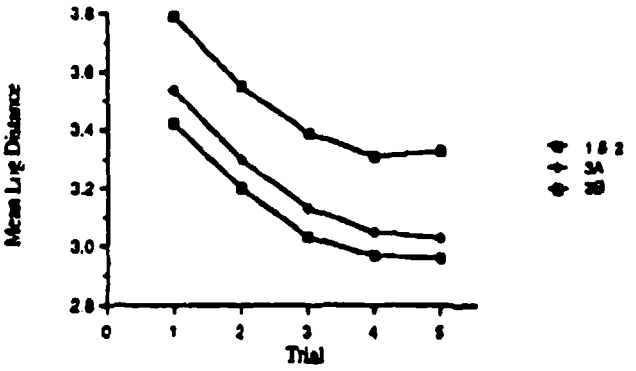


Figure 4-1. Mean log distance score by AFQT category trial

To determine if these practice effects were moderated by cognitive ability, this sample was also divided into three subgroups according to AFQT category. Four examinees belonged to Category 1 and were combined with 37 examinees belonging to Category 2. Of the remaining examinees for whom ASVAB scores were available, 39 belonged to Category 3A, and 31 belonged to Category 3B. Means and standard deviations of clipped mean decision time scores computed separately by category and test trial are reported in Table 4-8.

The results in Table 4-8 suggest that the clipped mean decision time scores, like the mean log distance scores reported above, differed systematically across the three AFQT category subgroups. However, between-subjects repeated measures ANOVA

results indicated that the between subjects effect of AFQT category was not statistically significant ($F(2,109)=1.99$, n.s.), nor was there a significant interaction between AFQT category and test trial ($F(8,436)= 1.04$, n.s.).

Target Identification: Median Movement Time. Means, standard deviations, and split-half (odd-even) reliabilities of median movement time scores for examinees with scores for each of the five trials are reported in Table 4-9. The results indicated that both the mean and the standard deviation of these scores decreased between the first and second trials before becoming relatively stable. The split-half reliability remained high and relatively unchanged (r_{xx} = .97-.98) across all five trials.

Repeated measures ANOVA results indicated that the differences in mean median movement time scores across trials were statistically significant ($F(4,452)=7.64$, $p<.001$). Results of paired-comparison t-tests comparing means of adjacent test trials, reported in Table 4-10, demonstrate that the only significant improvement in performance occurred between the first and second trials.

Means and standard deviations of median movement time scores computed separately by AFQT category and test trial are reported in Table 4-11. Results of the between-subjects repeated measures ANOVA indicated that neither the between-subjects effect of AFQT category ($F(2,108)=3.02$, n.s.) nor the interaction between AFQT category and test trial ($F(8,432)=0.92$, n.s.) was statistically significant.

Target Identification: Percent Correct. Means, standard deviations, and split-half (odd-even) reliabilities of percent correct scores for examinees with scores for each of the five trials are reported in Table 4-12. These results indicated that the mean and standard deviation of the percent correct scores were relatively stable across all five trials. Also, the reliability of these scores was lower and subject to greater fluctuation (r_{xx} = .69-.82) than that reported for the other measures.

Repeated measures ANOVA results indicated that the differences in mean percent correct scores across trials were not statistically significant ($F(4,452)=1.08$, n.s.).

Finally, means and standard deviations of percent correct scores computed separately by AFQT category and test trial are reported in Table 4-13. Results of between-subjects repeated measures ANOVA for these data indicated that neither the between-subjects effect of AFQT category ($F(2,108)=1.08$, n.s.), the within-subjects effect of test trial ($F(4,432)=1.14$, n.s.), nor the interaction between AFQT category and test trial ($F(8,432)=0.80$, n.s.) was statistically significant.

Table 4-8

Descriptive Statistics for Clipped Mean Decision Time Score by AFQT Category and Trial

AFQT Category (n)	Trial	Mean	SD
1 & 2 (41)	1	1.58	0.51
	2	1.20	0.38
	3	1.09	0.37
	4	1.02	0.37
	5	0.92	0.32
3A (39)	1	1.83	0.80
	2	1.40	0.63
	3	1.25	0.46
	4	1.10	0.41
	5	1.00	0.26
3B (31)	1	1.78	0.52
	2	1.45	0.54
	3	1.30	0.47
	4	1.16	0.41
	5	1.08	0.38

Table 4-9

Descriptive Statistics for Median Movement Time Score by Trial (N=114)

Trial	Mean	SD	Odd-Even r_{xx}*
1	0.35	0.13	0.98
2	0.32	0.09	0.97
3	0.32	0.09	0.97
4	0.32	0.08	0.97
5	0.31	0.08	0.97

*corrected for length using Spearman-Brown

Table 4-10

Paired-Comparison t-Tests for Mean Change in Median Movement Time Score Between Adjacent Trials

Trials	Mean Change	SD	t	p-value
1 and 2	0.03	0.01	2.78	0.0063
2 and 3	0.00	0.01	-0.24	0.8091
3 and 4	0.01	0.01	1.50	0.1375
4 and 5	0.00	0.00	0.95	0.3450

Table 4-11

Descriptive Statistics for Median Movement Time Score by AFQT Category and Trial

AFQT Category (n)	Trial	Mean	SD
1 & 2 (41)	1	0.32	0.05
	2	0.30	0.06
	3	0.30	0.06
	4	0.30	0.07
	5	0.30	0.07
3A (39)	1	0.39	0.16
	2	0.34	0.09
	3	0.35	0.12
	4	0.33	0.09
	5	0.33	0.10
3B (31)	1	0.36	0.14
	2	0.33	0.10
	3	0.33	0.10
	4	0.32	0.08
	5	0.31	0.08

Table 4-12

Descriptive Statistics for Percent Correct Score by Trial (N=114)

Trial	Mean	SD	Odd-Even r_{xx}*
1	0.93	0.09	0.82
2	0.94	0.07	0.70
3	0.94	0.07	0.69
4	0.95	0.07	0.72
5	0.94	0.08	0.82

*corrected for length using Spearman-Brown

Table 4-13

Descriptive Statistics for Percent Correct Score by AFQT Category and Trial

AFQT Category (n)	Trial	Mean	SD
1 & 2 (41)	1	0.95	0.04
	2	0.96	0.04
	3	0.94	0.06
	4	0.96	0.05
	5	0.94	0.10
3A (39)	1	0.92	0.11
	2	0.93	0.10
	3	0.93	0.08
	4	0.94	0.09
	5	0.93	0.10
3B (31)	1	0.93	0.09
	2	0.94	0.05
	3	0.95	0.05
	4	0.95	0.05
	5	0.95	0.04

Discussion

The analyses reported above clearly demonstrate that performance on computerized psychomotor and perceptual speed tests can be substantially influenced by practice. Regarding the Two-Hand Tracking test, the mean log distance scores decreased significantly across four successive administrations of this test prior to leveling off during the fifth and final administration. In comparison to mean log distance scores obtained during initial testing, scores obtained during the second test administration improved (on average) by approximately one-half of a standard deviation, and those obtained during the fourth administration improved by more than a full standard deviation. These results were not moderated by the AFQT category of the examinee.

For the Target Identification test, the results indicated that the influence of practice on test performance depended on which aspect of performance was considered. Whereas clipped mean decision time scores decreased significantly across each of the five test administrations (over one standard deviation between the first and fifth trials), median movement time scores demonstrated significant improvement only between the first and second trials (approximately one-fourth of a standard deviation), and percent correct scores showed no improvement at all. None of these results was moderated by AFQT category.

The implications of these findings for the operational use of these tests depend on the degree to which the practice effects reported in this investigation generalize to circumstances likely to be encountered in non-research settings. For instance, in the present research effort, intervals of only one minute separated each administration of the test. In "real life," however, it is unlikely that examinees would be able to practice immediately prior to operational testing. A worthwhile future research project would be to vary the amount of time between practice sessions to determine the rate at which these practice effects decay over time.

It is relevant to note that prior research conducted using 479 examinees from the concurrent validation (CV) sample of Project A found that mean log distance scores for Two-Hand Tracking improved one-fourth of a standard deviation between two trials separated by a one-month interval (Toquam, Peterson, Rosse, Ashworth, Hanson, & Hallam, 1986). That improvement was only half of that found between the first two trials in the present research. Likewise, improvements on Target Identification were also smaller over the longer interval than those reported in the present research. In particular, whereas decision time scores improved by approximately one-third of a standard deviation over the one-month interval between the two test trials administered to the CV sample, they improved

approximately three-fifths of a standard deviation between the first two trials of the investigation reported here.

It would also be informative to determine the extent to which the effects are attributable to the similarity between the items and equipment used during practice and those used with the actual test. Practice in the present investigation consisted of exactly the same test items and computer equipment (software and hardware) as the "operational" test. It is unlikely, however, that examinees in real life would have access to either of these prior to operational testing.

If the present results <u>are</u> found to generalize to more realistic circumstances, it might become necessary to ensure that all examinees taking the test have the same amount of practice prior to operational testing. Otherwise, examinees without practice would be at an unfair disadvantage. This would be particularly problematic if examinees from protected subgroups were less likely to have access to practice than other examinees.

Ensuring equal practice among examinees would probably necessitate administering sufficient numbers of practice items (prior to the operational items), such that any additional items would no longer result in improved performance. In the present investigation, performance on the Two-Hand Tracking test failed to improve after the fourth test administration (6 practice items + 18 test items, four times each = 78 items). However, it is possible that this was only a temporary stabilization in the learning curve for this test, and performance might improve further on subsequent trials. For the Target Identification test, clipped mean decision time scores improved throughout all five trials. Clearly, increasing the number of practice items to remove differential practice effects across examinees would lengthen both of these tests substantially.

Besides increasing the time required to administer the tests, the administration of additional practice items could also have an impact on their validities. Results from skill acquisition research (e.g., Ackerman, 1988) suggest that the abilities required for performance on various tasks change with practice. Thus, it is possible that the Two-Hand Tracking test is a purer measure of psychomotor ability (and less of a measure of cognitive ability) when it is scored following the administration of a series of practice items. If so, the incremental validity associated with this test (when used in conjunction with the ASVAB) may actually increase with practice. Similarly, the clipped mean decision time score of the Target Identification test may become a better measure of perceptual speed as experience is gained on that test. To date, as indicated earlier, both tests have only been validated with naive examinees. Therefore, whether the validities of these tests

would actually increase, decrease, or stay the same with practice is yet another question for future research.

It should be pointed out that the occurrence of practice effects is by no means limited to computerized or apparatus-based psychomotor and perceptual speed tests. For example, the United States Employment Service (U.S. Department of Labor, 1970) reports one-week test-retest improvements on the cognitive aptitude subscores of the General Aptitude Test Battery (GATB) ranging from one-fifth to one-quarter of a standard deviation. The USES also reports two-week improvements on the perceptual aptitude subscores ranging from one-half to four-fifths of a standard deviation, and three-week improvements on the Mark Making subtest (a paper-and-pencil psychomotor test) ranging from two- to four-fifths of a standard deviation.

On the other hand, improvements in test performance on the subtests of the Armed Services Vocational Aptitude Battery (ASVAB) reported by Christal (1989) are not nearly as large as those reported above. In fact, the largest improvement between initial and follow-up testing was only one-twelfth of a standard deviation (for the Coding Speed subtest), with most subtests demonstrating no improvements at all. Perhaps the lack of practice effects in Christal's study can be attributed somewhat to the longer duration between test sessions (up to six months) in comparison to the very short durations (one minute) used in the present research or the one- to three-week durations used in studying the GATB. Also, it may be that examinees are generally more familiar with the multiple choice format or content domain of many of the ASVAB's subtests (e.g., paragraph comprehension, word knowledge, math knowledge).

CHAPTER 5

BETWEEN-TEST ORDER EFFECTS ON COMPUTERIZED TESTS OF PERCEPTUAL SPEED AND PSYCHOMOTOR ABILITY

In this chapter, we examine the effects of test administration order on three computerized tests: One-Hand Tracking, Two-Hand Tracking, and Target Identification.

Previous research has shown these tests (in particular, Two-hand Tracking) to have incremental validity over ASVAB in predicting measures of gunnery performance (e.g., Graham, 1988; Smith & Walker, 1988; Busciglio, Silva, & Walker, 1990). However, in the versions of the computerized battery with which these results were obtained, the tests were always administered in the same order. Specifically, One-Hand Tracking was administered first, Two-Hand Tracking second, and Target Identification third.

As described throughout this report, the two tracking tests are very similar to one another and use the same kinds of items. Therefore, it is possible that One-Hand Tracking serves as a warmup, or even a practice session, for Two-Hand Tracking. If so, it may be that the validity results for Two-Hand Tracking are dependent on that test being administered after One-Hand Tracking. That is, it is possible that Two-Hand Tracking could have less (or perhaps even greater) validity if it were to be administered either by itself or following One-Hand Tracking. Likewise, the validity of One-Hand Tracking might be different if it were to be administered after Two-Hand Tracking. The purpose of the present research was to determine if the scores on these tests (including Target Identification) change as a function of their sequence of administration.

Sample

Data were collected at the Fort Knox Reception Battalion from 806 examinees with no previous military experience. Each examinee was assigned to one of the following four test administration order conditions:

1. One-hand Tracking, Two-hand Tracking, Target Identification (n=199);

2. Two-hand Tracking, One-hand Tracking, Target Identification (n=202);

3. One-hand Tracking, Target Identification, Two-hand Tracking (n=200); and

4. Two-hand Tracking, Target Identification, One-hand Tracking (n=205).

Armed Services Vocational Aptitude Battery (ASVAB) scores were available for 796 of these examinees. The mean Armed Forces Qualification Test (AFQT) score was 58.59, and the standard deviation was 17.94.

Measures

All examinees received each of the three computerized tests in one of the four orders listed above. Detailed scoring procedures for the three tests are described in detail in Chapter 2.

Results

One- and Two-Hand Tracking. Means, standard deviations, and split-half (odd-even) reliabilities of the mean log distance scores for the two tracking tests are reported by condition in Table 5-1. Analysis of variance results indicate that the effect of test administration order on mean log distance scores was not significant for either One-Hand Tracking ($F(3, 801)=0.40$, n.s.) or Two-Hand Tracking ($F(3, 802)=0.81$, n.s.). Inspection of the results in Table 5-1 suggests, however, that the variance of the One-Hand Tracking mean log distance scores was not equal across the four conditions. Results of F-tests for homogeneity of variances indicate that the variance of One-Hand Tracking scores associated with the first and third test administration orders (in which One-Hand Tracking preceded Two-Hand Tracking) were significantly smaller than the variances associated with the second and fourth test administration orders (in which Two-Hand Tracking preceded One-Hand Tracking). The results of these comparisons are reported in Table 5-2. Also, the reliabilities of the scores for both tests were consistently high ($r_{xx}=.96$ or greater) regardless of test administration order.

Target Identification. Means, standard deviations, and split-half (odd-even) reliabilities for the three Target Identification scores are reported in Table 5-3. Analysis of variance results indicate that two of the three scores varied significantly across the four conditions. Specifically, the effect of order was statistically significant for the clipped mean decision time scores ($F(3, 798)=2.68$, $p<.05$) and the percent correct scores ($F(3, 798)=2.97$, $p<.05$), but not for the median movement time scores ($F(3, 798)=1.06$, n.s.). Note, however, that results of F-tests of homogeneity of variance indicated that the variance of the median movement time scores for the third test administration order was significantly greater than that associated with each of the other three conditions ($p<.05$).

Table 5-1

Descriptive Statistics for One-Hand Tracking and Two-Hand Tracking Mean Log Distance Scores by Test Administration Order

Test Score	n	Order	Mean	SD	Odd-Even r_{xx}*
Mean Log Dist. (One-Hand)	198	1 (H1,H2,TI)	2.73	0.29	0.96
	202	2 (H2,H1,TI)	2.71	0.36	0.98
	200	3 (H1,TI,H2)	2.75	0.27	0.96
	205	4 (H2,TI,H1)	2.72	0.38	0.98
Mean Log Dist. (Two-Hand)	199	1 (H1,H2,TI)	3.54	0.44	0.97
	202	2 (H2,H1,TI)	3.59	0.42	0.97
	200	3 (H1,TI,H2)	3.60	0.44	0.97
	205	4 (H2,TI,H1)	3.60	0.41	0.97

*corrected for length using Spearman-Brown

Note: H1 = One-Hand Tracking; H2 = Two-Hand Tracking; TI = Target Identification.

Table 5-2

Results of F-Tests Comparing Variances of One-Hand Tracking Mean Log Distance Scores from Different Test Administration Orders

Test Orders Compared	df_{num}	df_{denom}	F-ratio	p
2 vs. 1	201	197	1.53	<.01
4 vs. 1	204	197	1.72	<.01
2 vs. 3	201	199	1.78	<.01
4 vs. 3	204	199	2.00	<.01

Table 5-3

Descriptive Statistics for Target Identification Scores by Test Administration Order

Test Score	n	Order	Mean	SD	Odd-Even r_{xx}*
Clipped Mean Decision Time	198	1 (H1,H2,TI)	1.46	0.50	0.97
	201	2 (H2,H1,TI)	1.53	0.55	0.97
	199	3 (H1,TI,H2)	1.61	0.58	0.97
	204	4 (H2,TI,H1)	1.57	0.51	0.98
Median Movement Time	198	1 (H1,H2,TI)	0.31	0.14	0.98
	201	2 (H2,H1,TI)	0.32	0.13	0.97
	199	3 (H1,TI,H2)	0.33	0.18	0.97
	204	4 (H2,TI,H1)	0.32	0.11	0.96
Percent Correct	198	1 (H1,H2,TI)	0.89	0.10	0.72
	201	2 (H2,H1,TI)	0.89	0.10	0.79
	199	3 (H1,TI,H2)	0.91	0.09	0.79
	204	4 (H2,TI,H1)	0.91	0.09	0.79

*corrected for length using Spearman-Brown

Note: H1 = One-Hand Tracking; H2 = Two-Hand Tracking; TI = Target Identification.

Table 5-3 shows that the clipped mean decision times achieved when Target Identification was administered after both tracking tests were slightly lower (i.e., faster) than when Target Identification was administered following only one of the tracking tests. Altogether, there are four possible comparisons between these two sets of conditions (Orders 1 and 2, respectively, vs. Orders 3 and 4, respectively). Of these, paired-comparison t-tests indicated that only two were statistically significant at the $p<.05$ level (Order 1 vs. Order 3 and Order 1 vs. Order 4). On the other hand, for the percent correct scores, performance was slightly better when Target Identification was administered after only one of the two tracking tests, rather than after both of them. These results also were significant ($p<.05$) for only two out of the four possible comparisons (Order 1 vs. Order 3 and Order 2 vs. Order 3).

Discussion

The analyses reported in this chapter indicate that the order of test administration had very little influence on the mean level of performance for each of the two tracking tests, but that the effects on two of the three Target Identification scores (clipped mean decision time and percent correct) were statistically significant. Although the magnitude of the effects on these latter scores tended to be small (on average, approximately two-tenths of a standard deviation), such results caution against administering these tests (at least Target Identification) in different orders across examinees. Additionally, the variability of mean log distance scores for One-Hand Tracking was significantly increased when that test was administered following Two-Hand Tracking. Although the explanation for this latter finding is unclear, the fact that the variance of these scores did vary with order of test administration suggests the possibility that the validity of the test could also vary with administration order. Unfortunately, examining the influence of test administration order on the validity of these tests was beyond the scope of the present effort.

CHAPTER 6

DEVELOPING THE SPATIAL ITEM POOL: ITEM DEVELOPMENT, SCREENING, AND CALIBRATION

One of the goals of the CMOS Project was to conduct research to upgrade Project A tests to justify their inclusion in the ECAT testbed. One part of this research involved enlarging the item pool of the Assembling Objects test. At the start of the project, there were 36 Assembling Objects items used in Project A. After considering the relative advantages of having a large number of items versus having more data per item from which to estimate item parameters, ARI developed an additional 144 items. In this project, data were collected on these new items, as well as on the 36 original items. This chapter describes the item and booklet development, as well as the procedures used for collecting the data. Also presented are recommendations for analyzing these data and constructing new Assembling Objects tests.

Item Development

The Assembling Objects test consists of two types of items: "puzzle" items and "tinker-toy" items. The stem of each "puzzle" item is a set of separated jigsaw-puzzle-like pieces. The examinee is presented with four response options, each of which consists of the same geometric shape, but which has been constructed from different sets of puzzle pieces. The examinee must determine which option was constructed with the same set of pieces as those in the stem.

"Tinker-toy" items require that the examinee fit together a picture of labelled parts by matching the letters indicating where the parts should touch. The stem is like an exploded view of machinery, and the four choices are different assemblies of the parts. A more detailed description of both types of Assembling Objects items can be found in Peterson (1987).

As indicated above, ARI created draft copies of 144 new Assembling Objects items. Half of these were "puzzle" items, and the other half were "tinker-toy" items. The process of reconstructing and amplifying the test specifications for Assembling Objects, and of creating the new items, is documented in Busciglio, Palmer, King, and Walker (1992, in preparation). The draft items were then given to AIR, which created camera-ready versions of the items using CORELDRAW software on a 486 IBM-compatible PC equipped with a high-resolution monitor. The initial drawings were reviewed by ARI, and revisions were made according to their comments. The original 36 items from Project A were also redrawn.

Booklet Development

The Assembling Objects Pilot Calibration was designed with four objectives in mind, the most important of which was to collect data with which to assess the quality of the new items. With 144 new items to evaluate, it was considered infeasible to collect sufficient data per item to calibrate them using analytic techniques based on item response theory (IRT). Instead, a sample target of 200 examinees per new item was sought in order to calculate classical item statistics.

The second objective of the Pilot Calibration was to collect additional data on the 36 original items. The smaller number of original items (36) made it possible to consider collecting data from an adequate sample to calculate IRT item parameter estimates. A sample of 1,200 examinees for the 36 original items was sought in order to allow the calculation of IRT estimates under the conditions used in the present data collection.

The third objective was to insure the administration of the new test booklets under maximum power conditions. (A pure power test is generally defined as a test in which every examinee is given all the time needed to attempt every test item. Almost no tests meet this criterion. Operationally, most experts accept as a power test a test for which 90% of the examinees have sufficient time to complete all of the items [Hartigan & Wigdor, 1989, pp. 103].) The original 36-item test booklet had been administered under somewhat less than maximum power conditions (12 minutes in which to complete 36 items) as part of the Project A longitudinal validation data collection. In that data collection, the average examinee completed 94% of the items, with the standard deviation of percent items completed equal to 14%. Thus, although the Assembling Objects test was certainly not highly speeded, many of the examinees did not complete all of the items. In order to ensure administration under power conditions, each of the developmental test booklets consisted of 48 items, and 50 minutes were allowed for the completion of all items. This allowed more than one minute per item, compared to 20 seconds per item for the administration conditions in the LV sample.

A fourth objective was to collect data with which to evaluate the influence of item administration order on item characteristics. For example, we would like to determine whether performance on "puzzle" items differs as a function of whether or not examinees have previously received "tinker-toy" items. The answer to this question (and others like it) has important implications for the calibration of the items and, ultimately, the development of operational alternate test forms. Therefore, as described below, the order of administration of the original items was systematically varied across test booklets.

Altogether, 15 test booklets were created. The contents of each booklet are summarized in Table 6-1. The original 18 "tinker-toy" items are identified as T1-T18, and the 18 original "puzzle" items are identified as P1-P18. Similarly, new "tinker-toy" items are labeled from T19 to T90; and new "puzzle" items are labeled from P19 to P90. As indicated in Table 6-1, odd-numbered booklets started with a block of 24 "tinker-toy" items, followed by a block of 24 "puzzle" items. Even-numbered booklets started with a block of 24 "puzzle" items, followed by a block of 24 "tinker-toy" items. The first 12 booklets consisted of the 36 original items, plus 12 new items (six of each type). For these 12 booklets, the last six items in each block were new items. Booklets 13, 14, and 15 contained only new items (24 of each type). Table 6-1 indicates that each new item was administered in two different booklets: once in one of Booklets 1-12, and once in one of Booklets 13-15.

Table 6-1

Summary of Test Booklet Item Content

	1st Item Block		2nd Item Block	
Booklet Number	Old Items	New Items	Old Items	New Items
1	T1-T18	T19-T24	P1-P18	P19-P24
2	P1-P18	P25-P30	T1-T18	T25-T30
3	T1-T18	T31-T36	P1-P18	P31-P36
4	P1-P18	P37-P42	T1-T18	T37-T42
5	T1-T18	T43-T48	P1-P18	P43-P48
6	P1-P18	P49-P54	T1-T18	T49-T54
7	T1-T18	T55-T60	P1-P18	P55-P60
8	P1-P18	P61-P66	T1-T18	T61-T66
9	T1-T18	T67-T72	P1-P18	P67-P72
10	P1-P18	P73-P78	T1-T18	T73-T78
11	T1-T18	T79-T84	P1-P18	P79-P84
12	P1-P18	P85-P90	T1-T18	T85-T90
13	----	T19-T42	----	P19-P42
14	----	P43-P66	----	T43-T66
15	----	T67-T90	----	P67-P90

Data Collection

Data were collected beginning in March 1991 and ending in May 1991 from a total of 1,561 Army recruits being in-processed at the Fort Leonard Wood, Missouri, Reception Battalion. Twelve recruits are not included in the discussion below because of missing data, leaving 1,549 examinees in the sample.

Booklets were administered in a spiral order to groups of examinees. The number of examinees who received each test booklet is reported in Table 6-2. This number ranged from 91 to 111. Altogether, 1,253 examinees received the original items. The number of examinees who received each of the new items ranged from 191 to 217.

Table 6-2

Number of Examinees Administered Each Test Booklet

Booklet Number	Number of Examinees
1	109
2	109
3	111
4	104
5	100
6	107
7	102
8	100
9	103
10	108
11	100
12	100
13	106
14	99
15	91

Analysis Recommendations

The remaining steps to be taken in developing alternative operational forms of the Assembling Objects test are to: 1) conduct psychometric analyses of the original and new items using data collected in this project; 2) use the results of those analyses to construct preliminary alternative test forms; and 3) examine the degree to which the preliminary alternate forms are parallel to the original test using data collected from an

independent sample. This section contains recommendations regarding each of these steps.

Conduct Item Analyses. The analyses in this step should begin with an explicit evaluation of the extent to which power conditions were met. This can be accomplished by examining the number of item omissions per item position. If the number of omissions increases noticeably for the later item positions, this would indicate that maximum power conditions were not met. In that case, the item statistics for items administered later in the test booklets should be interpreted with caution.

Next, item statistics should be computed. These item statistics should include the proportion of examinees correctly answering each item and the point-biserial correlation between each item and the total test score. For the new items, the total test score should be computed as the number of original items answered correctly. For the original items, the total score should be computed as the number of original items answered correctly not including the item being correlated with the total. In addition, for both the original and new items, information should be gathered on the proportion of examinees selecting each option. Further review of an item should occur if: 1) the item does not discriminate between examinees with high and low scores on the original test form; 2) the item score variance is too small (i.e., everybody tends to give the same response); or 3) too many examinees omit the item.

Factor analyses of the item scores could also be carried out to confirm the a priori dimensionality of the items. The dimensions should be inferred, if possible, from Busciglio, Palmer, King, and Walker's (1992, in preparation) description of the test specifications. That is, items written to the same specification should be hypothesized to load on a dimension different from items written to different specifications. If the a priori structure is not confirmed, exploratory factor analyses could be conducted to further investigate the structure of the original and new items. These analyses are important for evaluating the extent to which the original and new items are measuring the same construct.

To examine the original items, IRT item parameter estimates can be calculated in addition to the above classical item statistics. To examine whether administering the Assembling Objects test under power conditions is different from "partially speeded" conditions, the obtained item statistics should be compared to the item statistics of the Project A LV sample. Also, to examine whether administering the "tinker-toy" or the "puzzle" items first makes a difference, the item statistics of the even and odd numbered booklets should be calculated separately and compared.

Construct Preliminary Alternate Test Forms. Ultimately, several alternate forms of the original paper-and-pencil subtest can be constructed from the total item pool. In the prior step, the new items should have been screened for item difficulty and item discriminability. Items passing these screens will be eligible for inclusion in the new test forms. The next step in making new test forms is to match items within each of the constructs used in the factor analysis in terms of item difficulty and item discrimination. This step would be followed by pairing one new item within each of these sets with each original item. The new items selected in this manner would constitute a new alternative test form, approximately parallel to the original form. Additional forms could be created using the same selection rule.

Preliminary analyses at the test score level should then be conducted on the original and preliminary alternate test forms. These analyses should include the computation of test score distributions (means, standard deviations) for each of the alternate forms, and the comparison of these distributions with that of the original test. Additional analyses that might be conducted at this step include the computation and comparison of internal consistency reliabilities, and the use of confirmatory factor analysis to compare the factor structures of the items on the alternate forms with the structures of the items on the original test.

Evaluate Preliminary Alternate Test Forms Using an Independent Sample. In some respects, recommendations for this step depend on the results of the preceding steps. If, for instance, one or more alternate forms are constructed which appear in all respects (mean, standard deviation, reliability, etc.) to be equivalent to the original test, it might be considered appropriate to conduct a "pre-operational" data collection, in which data are collected on the new and original tests from fairly large samples, using operational testing conditions (time limits, etc.). Analyses would then be conducted to confirm that the tests are indeed parallel; or, if necessary, techniques could be used to equate the different forms.

On the other hand, if the evidence for the existence of parallel forms is not as strong, it may be more appropriate to collect additional data using the same procedures as were used in the present effort. Prior to this future data collection, items which were identified as having problems could be modified, and new items could perhaps be written to fill voids identified in the preceding analyses. The analyses from this future new data collection should proceed in the same manner as described in the previous steps. That is, item analyses and "matching" should be conducted, followed by the construction and evaluation of new preliminary test forms (or the re-examination of the previous created forms). As before, test score distributions and

reliability estimates for the alternate forms should be compared with the original test, as should the factor structures of the test items.

CHAPTER 7

SUMMARY AND DISCUSSION

In this chapter we summarize and discuss the findings from the research described in the previous chapters and make suggestions for next steps. The chapter is organized around the three major classes of measures that we have examined: the spatial tests, the ABLE, and the psychomotor/perceptual speed tests.

Spatial Tests

Results in Chapter 2 showed that modifying the Project A instructions (as they had been implemented on the ECAT battery) to make them as simple and easy to read as possible did not cause any discernible improvements in the characteristics of the three spatial tests. There were no significant differences between the two versions of the instructions in terms of time to read the instructions, time to complete the tests, observed mean scores, internal consistency reliabilities of test scores, or correlations of AFQT with test scores or time to read the instructions. There was a statistically significant difference for Spatial Reasoning with regard to perceived clarity of instructions, but this was not a practically significant difference, and the other two tests showed no significant differences. In retrospect, this lack of further improvement might have been expected given the fairly extensive, iterative process of development of these tests over the course of Project A.

Comparisons of the paper-and-pencil and computer-administered versions of the spatial tests, described in Chapter 3, showed significant differences in mean observed scores for Assembling Objects and Spatial Reasoning (in favor of the paper-and-pencil versions). In all other respects, the three tests seemed very similar across the two testing methods (internal consistency reliabilities, correlations among the tests, and correlations with AFQT showed no differences). Differences in time allowed to take the tests seemed to account for some of the score differences for Assembling Objects, but not for Spatial Reasoning. Use of item difficulty statistics to identify and examine those items that were more difficult on the computer led to the conclusion that the more visually complex items with more small pieces were more difficult on the computer, perhaps because of the poorer resolution available on the computer versus the paper-and-pencil version.

In Chapter 6 we described the development of new items for the Assembling Objects test, and the collection of data on these

new items and the original items. We also made recommendations for analyses of these data.

Where should research go from here? Obviously, analyses described in Chapter 6 of the new Assembling Objects items should proceed. Further research on instructions for these tests does not seem productive; they seem excellent as they are. In general, both the paper-and-pencil and computer-administered versions of the tests have adequate psychometric properties, so no drastic revisions are in order for either version of the tests.

If computer-administered versions of these tests are to be eventually used, however, additional efforts should be undertaken to explain and/or control the observed score differences between the paper-and-pencil and computer-administered versions. In this regard, time allowed to take the computer-administered versions of the tests should be reconsidered. The rationale for allowing less time for the computer-administered version of the Assembling Objects test (in comparison to the paper-and-pencil version) is obscure. Depending on the goal, the method of controlling or allowing time to complete each item and the total test will vary. Is the goal to achieve a pure power test in each case, or is the goal to maintain the same level of speededness across the two modes of administration? Is the goal to make the computer-administered versions the best possible tests of the construct without regard to maintaining comparability with the paper-and-pencil versions? Practically speaking, it would seem desirable to have the two modes of administration produce scores as comparable as possible, so that maximum flexibility of implementation is achieved, i.e., it should be a matter of indifference to the examinee which mode of the test is given.

It may be that the score differences between the two versions will disappear when the computer-administered versions are installed on machines with higher resolution. This hypothesis gains support from analyses in Chapter 3 which seemed to indicate that the poorer resolution on the computer screen (compared to the paper-and-pencil version) increased the difficulty of Assembling Object items. Only empirical evidence can answer this question.

ABLE

Differences in order of presentation of ABLE items produced no interpretable effects on characteristics of the ABLE scale scores. Internal consistency reliabilities showed little differences across orders. Three of the fourteen ABLE scales (not including the Non-Random Response scale) showed mean scores that were statistically significantly higher for the original item order, but all of these differences were no more than a quarter of a standard deviation. Additionally, the percent of

items appearing early versus late for each of these scales were differently affected by changing from original to reversed item order. Yet the mean scale differences were in all the same direction. We concluded that item order does not appear to be a major factor with regard to ABLE scale score.

On the other hand, a significant and interpretable pattern of differences was found between the two methods of administering the ABLE. The differences between the computer-administered and paper-and-pencil versions strongly indicate that examinees may have taken the computer-administered version much more seriously and paid closer attention when responding. The computer-administered version resulted in many fewer examinees failing the random response screen and showed greater score variances on a majority of the substantive scales, greater internal consistency reliabilities, higher levels of correlation among the ABLE substantive scales, and a more finely differentiated factor structure. Importantly, only one scale showed a small, but statistically significant, mean score difference, indicating that tendencies to inflate or deflate scores were not associated with either mode of administration. If these findings hold up in other and larger data sets, then it seems encouraging news for the use of temperament/biodata inventories in a computer-administered format. Possibilities include shortening the inventory because of the increased reliability, and achieving better classification of examinees based on ABLE scores because of the cleaner differentiation of constructs.

Future research should focus on verifying and extending the results described just above. Does the pattern of results indicating closer attention by examinees hold up in more diverse and larger samples? Are there differences in validity coefficients for training, job performance, or attrition to buttress the other favorable, psychometric characteristics associated with computer administration? Could this increased carefulness on the part of the examinees be further enhanced by monitoring their time to respond to items and cautioning examinees that they are proceeding too slowly or too quickly? Or would such feedback be detrimental to the psychometric characteristics of the ABLE?

Psychomotor/Perceptual Speed Tests

The results reported in Chapter 2 indicated that changes to the Project A/ECAT instructions had even less influence on the characteristics of the psychomotor/perceptual speed tests than they did on the spatial tests. There were no significant differences between the two versions of the instructions with respect to time to read the instructions, time to take the test, observed mean scores, internal consistency reliability, or correlations of AFQT with test scores or time to read the

instructions. Nor were there any differences between the old and new instructions with respect to their perceived clarity.

Analyses in Chapter 4 demonstrated that performance on both Two-Hand Tracking and Target Identification was substantially influenced by practice. (Note that One-Hand Tracking was not included in this particular investigation.) For Two-Hand Tracking, mean log distance scores improved significantly across the first four (out of five) administrations of the test. Specifically, mean performance on the fourth administration was more than a full standard deviation better than mean performance on the first trial. Regarding Target Identification, improvements in test performance depended on the particular test score being considered (recall that three different scores are computed for this test). For example, clipped mean decision time scores improved by more than a full standard deviation between the first and fifth trials, but median movement times improved significantly only between the first and second trials (by approximately one-fourth of a standard deviation). None of these results was moderated by AFQT category.

Finally, results reported in Chapter 5 indicated that the order of test administration had very little influence on the mean level of performance for either of the two tracking tests, although the variability of mean log distance scores for One-Hand Tracking was significantly increased when that test was administered following Two-Hand Tracking. On the other hand, the effect of order of test administration was statistically significant for the mean scores of two of the three Target Identification scores (clipped mean decision time and percent correct). The magnitude of the effects on these latter scores tended to be small (approximately two-tenths of a standard deviation); however, such results caution against administering these tests (at least Target Identification) in different orders across examinees.

The existence of the sizeable practice effects for these tests is somewhat troublesome. If the tests were to be made operational and examinees had unequal access to opportunities to practice on tasks similar to these tests, then those with access might have an unfair advantage. However, as discussed at some length in Chapter 4, the size of the practice effects reported here are probably greater than would be experienced in the "real world." In these experiments, examinees practiced on exactly the same equipment used for testing and immediately prior to being tested. In the real world, these circumstances could only be approached, but not obtained. Nevertheless, future research could focus on two topics and their interaction: the effect of degree of similarity of practice task to the actual test and the effect of length of time between practice and actual testing. This research would provide information about the probable size

of score differences between practiced and unpracticed examinees, given different scenarios of feasible practice.

Another unknown concerning practice is its effect on validity. We know that scores on these tests, without practice, show validity for predicting job performance in a number of MOS and, in particular, for gunnery tasks (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Busciglio, Silva, and Walker, 1990). We do not know if practice increases, decreases, or has no effect on validity. If practice increases validity, then it would be beneficial to insure that all examinees had sufficient practice to provide scores with the highest validity. The research design to obtain the data to evaluate practice effects on validity is relatively straightforward, but is probably only feasible for gunnery training performance. For these MOS (11H and 19K), there would be large numbers of soldiers passing through training and providing criterion scores on job tasks most likely to be affected by practice. Different groups of soldiers could be given different amounts and types of practice (e.g. 50, 100, 150, or 200 practice items administered massed or spaced). If it turns out that practice has an effect on validity, then it probably does not bode well for using psychomotor tests in general testing for all applicants for enlistment. But it certainly could still have a place for testing after enlistment in order to identify the "best bets" for gunnery MOS. Opportunities to provide sufficient practice would exist in the latter case, but would probably not be feasible in the former case.

Concluding Remarks

We think the results of the CMOS research bear out the wisdom of carefully investigating the effects of variables likely to become salient as personnel tests more closer toward becoming operational. Some issues now seem less important or troublesome; others perhaps more so. There was even some good news.

However, we have only investigated the Project A tests that now appear most likely to become operational; others remain to be researched should they also move into operational consideration. Furthermore, as just discussed, not all the issues investigated were resolved and it is likely that issues not investigated may become important in the future.

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 228-318.

Bilodeau, E. A. (1952). Transfer of training between tasks differing in degree of physical restriction of imprecise responses. USAF Human Resources Research Center Research Bulletin, 52.

Busciglio, H. H., Palmer, D. K., King, I. H., & Walker, C. B. (1992, in preparation). Development of new forms of the Assembling Objects Test.

Busciglio, H. H., Silva, J. M., & Walker, C. B. (1990, June). The potential of new Army tasks to improve job performance. Paper presented at the Army Service Conference, Durham, N.C.

Campbell, J. P. (Ed.). (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year (ARI Technical Report 746). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A193 343)

Campbell, J. P. (Ed.). (1988). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year (ARI Technical Report 792). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A198 856)

Campbell, J. P. (1990a). An overview of the Army selection and classification project (Project A). Personnel Psychology, 43, 231-239.

Campbell, J. P. (Ed.). (1990b). Project A: The U.S. Army selection and classification project. Special issue of Personnel Psychology, 43, 231-378.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felker, D. B., De Vera, M. V., & Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. Personnel Psychology, 43, 277-300.

Campbell, J. P., & Zook, L. M. (Eds.). (1991). Improving the selection, classification and utilization of Army enlisted personnel: Final report on Project A (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A242 921)

Campbell, J. P., & Zook, L. M. (Eds.). (1992). Building and retaining the Career Force: New procedures for accessioning and assigning Army enlisted personnel: Annual report, 1990 fiscal year (ARI Technical Report 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A252 675)

Cascio, W. F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Christal, R. E. (1989). Estimating the contribution of experimental tests to the Armed Forces Vocational Aptitude Battery (AFHRL-TP-89-30). Brooks AFB, TX: Manpower and Personnel Division, U.S. Air Force Human Resources Laboratory.

Graham, S. E. (1988, December). Selecting soldiers for the Excellence in Armor Program. Paper presented at the annual meeting of the Military Testing Association, Arlington, VA.

Hays, W. L. (1973). Statistics for the social sciences. 2nd ed. New York: Holt, Rinehart and Winston.

Hartigan, J. A., & Wigdor, A. K. (1989). Fairness in employment testing. Washington, DC: National Academy Press.

Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 10, 193-206.

Johnston, W. B., & Packer, A. J. (1987). Workforce 2000. Indianapolis, IN: Hudson Institute.

Jones, M. B. (1969). Differential processes in acquisition. In E. A. Bilodeau & I. McD. Bilodeau (Eds.). Principles of skill acquisition (pp. 141-170). New York: Academic Press.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent square roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. Psychometrika, 41, 341-348.

Peterson, N. G. (Ed.). (1987). Development and field test of the trial battery for Project A: Final report (ARI Technical Report 739). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A184 575)

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. Personnel Psychology, 43, 247-276.

Peterson, N., Russell, T., Hallam, G., Hough, L., Owens-Kurtz, C., Gialluca, K., & Kerwyn, K. (In preparation). Chapter IV. Longitudinal validity predictor data analyses. In J. P. Campbell (Ed.), Building the Career Force: First Year Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Smith, E. P., & Walker, M. (1988, December). Testing psychomotor and spatial abilities to improve TOW gunner selection. Paper presented at the annual meetings of the Military Testing Association, Arlington, VA.

Toquam, J., Peterson, N. G., Rosse, R., Ashworth, S., Hanson, M. A., & Hallam, G. (1986, March). Concurrent validity data analyses: Cognitive paper-and-pencil and computer-administered predictors (trial battery). Paper presented at SAC Meeting.

U.S. Department of Labor (1970). Manual for the USES General Aptitude Test Battery. Section III: Development. Washington, DC: Manpower Administration, U.S. Department of Labor.

Walker, C. B. (1989). The process and results of predictor testing in TRADOC's Skills Selection and Sustainment (S3) Program (ARI Working Paper WP-RS-89-23). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wise, L. L., Wang, M., & Rossmeissl, P. (1983, June). Development and validation of Army selection and classification measures, Project A: Longitudinal Research Database Plan. Alexandria, VA: Human Resources Research Organization.

APPENDIX A

A COMPARISON OF ECAT AND PROJECT A TESTING HARDWARE AND SOFTWARE

Introduction

The initial efforts in the development of the computerized test battery for the Army Research Institute's (ARI's) Project A began in late 1982. Portable microcomputers were first becoming available for wide-spread use. The first pilot tests occurred on 15 January 1984 at a MEPS station in Minneapolis, Minnesota.

At that time, ARI chose to incorporate this new computer capability into the batteries of tests being developed and validated as a part of Project A. The final version of the computerized test component of the project was completed in late 1985. The computer software development was not performed as a separate effort from Project A. The computer programming was done by a psychologist, R. L. Rosse, Ph.D., who was a participating member of the Project A research team.

The test constructs targeted for computer administration included primarily those in which the psychometric properties of testing could be most enhanced by computer administration: eye-hand coordination, spatial reasoning, pattern recognition, and short-term memory. A response pedestal was developed to provide an appropriate means of response acquisition for these constructs. Its technology was based upon that used for home computer games.

The microcomputer initially chosen for pilot testing was offered by Compaq as an "IBM PC compatible" computer. It had just been placed on the market. After pilot and field testing, bids were solicited for provision of a large number of machines during the primary data collection. The machine that was selected was manufactured by the Seequa Corporation. Two slightly different versions of the Seequa machine were produced (the Chameleon Plus and the XL). These machines were functionally equivalent, and both were very similar to the developmental computer. A large number of the Seequa microcomputers were procured from Seequa Corporation.

Newer versions of the computerized tests were developed by the U. S. Navy. A panel of the Services' testing experts chose to incorporate selected tests from the computerized Project A battery into the Enhanced Computer Administered Tests (ECAT) Project. The Navy was using a different microcomputer known as the UNIX Hewlett-Packard Integral Personal Computer (HP-IPC) for the ECAT. The Project A testing software (as explained later)

could not be used with the Navy's HP-IPC. Because of this, a porting (software conversion) effort was required. This lead to a major effort contracted to RGI, Inc., which resulted in tests which were administered by new software on the newer microcomputer.

The objective of the ECAT implementation of the Project A tests was to preserve the psychological constructs and psychometric properties of the Project A tests on the HP-IPC platform. Thus, the conversion of the Project A tests for ECAT usage involved conversion and rewriting of the software and adaptation of the functions to a different hardware environment so as to obtain testing results as equivalent as possible to the original tests. What follows in this chapter is an attempt to characterize the known similarities and differences between the computerized tests developed in Project A and those which were developed as a part of the ECAT testing project.

Overview of Testing Environment

The Project A tests were designed to be administered by microcomputer in the predictor data gathering effort. To minimize intervention by a testing monitor, the testing was designed to provide examinees all necessary information on the computer screen. Control of the testing session was provided by the instructions presented by the computer throughout the testing session.

The test monitor had the responsibility of preparing the computers for test administration and of assuring that the stored data for the test subject was correctly identified at the completion of test administration. The reason for this design was that data gathering effort often involved many test subjects being processed in group testing sessions.

The ECAT testing environment is essentially the same except that the test administration protocol permitted the test subject to interrupt testing (by pressing a "HELP" button).

Comparisons of Computer Hardware used in Project A and ECAT

The Seequa computer used in the operational versions of the Project A data gathering is an early IBM PC compatible microcomputer. It was designed as an IBM PC compatible computer which is in a transportable package with a handle for carrying. The HP-IPC used by the Navy is a microcomputer designed to fill the same purpose. Both microcomputers have video displays built into the frame on the front-left, disk drives built into the front-right beside the display, detachable keyboards that also function as front covers to the display and disk drives, and both may be carried by a handle. The following subsections discuss relevant similarities and differences.

CPU and Storage. The Seequa uses an 8088 CPU released by Intel in the late 1970's. This is a 16 bit data/address processor that uses an 8 bit data bus. The processing speed of the 8088 CPU is 5 MHz. It has 2 floppy disk drives: a 360 K drive (designated as device A:) and a 1.2 MB drive (designated as device "C:"). The PC system performs all file input and output directly to and from floppy diskettes. The Seequa Chameleon computers that were used have 256K of RAM and 16K of ROM.

The HP-IPC uses the 68000 CPU released by Motorola in the early 1980's. This is a 32 bit data/address processor that uses a 24 bit data bus. The processing speed of the HP's CPU is 8 MHz. The HP-IPC also incorporates a graphics co-processor which performs graphics operations that must be performed by the CPU in the Seequa Chameleon. The design of the HP-IPC results in a considerable speed and performance advantage over the Seequa computer.

Display Characteristics. The video displays constitute an important difference between the HP-IPC and the Seequa Chameleon. The Seequa's screen is a 6½ inch square, green/black, composite cathode ray tube (CRT). Its oval shaped pixels blur at the edges so as to soften text and graphics. The HP-IPC's screen is an amber, 8 x 4 inch ElectroLuminescent polarized display. Its pixels are square and distinctly sharp.

A standard Color Graphics Adaptor (CGA) was used in the Seequa Chameleon (although, the color options of the CGA card were not used in the Project A tests). The CGA provides for two modes of graphics resolution: (1) low resolution, and (2) high resolution. Only the high resolution graphics mode was used in the Project A tests. The high resolution graphics provided a monochrome display of 640 by 200 pixels. In this mode, the oval pixels measure about 2.27 times taller than they do in width and are blurred on the display so as to blend into fairly continuous lines.

The graphics display for the HP-IPC screen is 512 by 255 pixels. The pixels of the display are square and sharply displayed so that no blending occurs. Vector images are perceived to look more "jagged."

The CGA card used in the Seequa provides two text modes: (1) 40 columns by 25 lines, or (2) 80 columns by 25 lines. Only the second mode, 80 x 25, was used in the Project A tests. Also, only the one font and character spacing that was supplied with the system was used. When textual information was presented in the Project A tests, the display was switched to textual mode if needed so that text was not displayed as a part of graphical displays. Even though text can also be displayed in the graphics mode, very little use was made of this capability in the Project A tests. (This was because the textual mode provided better

visual presentation of text and better control of the process of presentation.) Minor exceptions to this included the word, "HIT" or "MISS" in the test involving acquisition and "shooting" of a target.

The HP-IPC has a graphics-based display, so the text size depends on the font in use. The standard font used in the ported ECAT software is the ROM 7x11 THIN set. This font allows a 72 column by 23 line display. The graphics-based display would not require the sharp distinction between displays of textual and graphical content that was required on the Seequa computers.

Peripheral Devices and Expansion Slots. The Seequa is configured as a conventional IBM-PC or clone. The operational version of the Seequa includes a detachable keyboard, a parallel adapter, and a serial communications adapter, none of which were used in actual testing. The keyboard attaches to the front of the PC.

The Seequa has five expansion slots into which expansion circuity may be inserted. These expansion circuits communicate with software through input/output (I/O) ports. Two of the expansion slots are used for the CGA card and the game board which was used with the Project A response pedestal for input of responses.

I/O ports on the HP-IPC include 2 generalized Human Interface Links (HIL) in the front of the computer, a standard IEEE (488.1) HPIB parallel port and 2 expansion slots in the back of the computer. The HIL ports accommodate a keyboard, mouse, trackball, or graphics tablet. Hard drives may connect to the HPIB port. Since many HP's are used for both the ECAT and the Accelerated CAT-ASVAB Project (ACAP) testing, an HP used for ACAP or ECAT typically includes a 1 MB memory board and an HP-IL board.

Operating Systems. The Seequa Chameleon uses Microsoft's DOS (version 2.11) operating system loaded from a diskette. DOS occupies about 50K of the 256K RAM. The Seequa version of the DOS system includes a Real Time Clock driver (RTC version 1.01).

The HP-IPC has an HP-UX Operating System (OS) built into ROM. HP-UX Release 5.0.1 is based on AT&T's UNIX System V.2.

Pedestal Hardware. The response pedestals used for both the Project A and ECAT tests are quite similar in external appearance and function but not identical. The Project A response pedestal is a black metal box with a footprint of 20 x 10 inches, and 3 inches high. The HPIB pedestal is a similar black metal box with dimensions of 19 x 10 x 3 inches.

Both pedestals contain the same pattern of colored buttons, a left-hand joystick, a right-hand joystick, a vertical slide, and a horizontal slide. The joysticks and slides are used to move a cursor (usually a crosshair) around the screen. Examinees use three of the buttons (BLUE, WHITE, YELLOW) for multiple choice selections. They use two RED buttons (LEFT and RIGHT) for prompts such as "Press a RED button to continue", or "Press a RED button to fire".

The testing software uses the four GREEN HOME buttons. The test subject is said to be in "home" position when the green buttons are depressed. Most of the test software was designed to require "home" position in order to initiate the presentation of test items.

On the Project A's response pedestal, a rotary dial is used for entering Social Security Numbers and responding to biographical questions. The rotary dial was omitted from the HBIP response pedestal but no effect upon test performance should be expected because the rotary dial was not used for test purposes. Also, a "HELP" key was added to the HPIB response pedestal and was intended to permit examinees to interrupt and temporarily suspend a test while receiving assistance from a test monitor.

The 20-inch width of the Project A response pedestal was considered an important issue in its design. Two variables of the test results were (1) the length of time from stimulus onset to release of the "home" position (either side), and (2) the length of time from release to pressing the response. The 20 inches were needed to space the "home" position (green buttons) sufficiently far from the response keys that the test subject could not easily reach both sets of keys without moving his or her hand.

The HPIB pedestal was made smaller (one inch more narrow). According to RGI, this was done so that it would fit inside the HP's computer box. Unfortunately, this sacrificed the spacing of the "home" position from the response buttons. This could affect the data collected from testing since the required movements for registering a response are somewhat different.

Response Pedestal Controller Board. The Project A pedestal uses a modified version of an "off-the-shelf" IBM-game board inserted into an expansion slot of the Seequa Chameleon computer. The method used for analog input from the joystick, sliding adjuster, or rotary dial is the same as the IBM-game board. The modification, fabricated by Seequa Corporation, was that the original IBM-game board provided for only four buttons and only four analog inputs (i.e., two joysticks). The modified board provides for double the number of inputs, i.e., 8 analog inputs and 8 buttons. This accommodates all of the controls on the

response pedestal: seven buttons, two joysticks, two sliding adjusters, and the rotary dial.

The functions of the controller board in the ECAT implementation are accomplished within the pedestal itself. Communication with the ECAT software is accomplished with a standard IEEE 488 Interface Bus. Hewlett-Packard's implementation of this bus is called the HPIB. The software to control devices connected to the HPIB is called a device driver. The HPIB pedestal is controlled by such a driver, which is a separate software package that is "loaded" into RAM before using the pedestal. ECAT loads the device driver automatically during its bootup phase. ECAT response pedestals require their own power supply.

Software Development--Porting Effort

The Pascal programming language was chosen for the development of the Project A tests on the basis of its relative simplicity, readability, and its "self-documenting" syntactical style. The Project A programmer used two early versions of Microsoft Pascal which were distributed during 1982-1984. These early versions implemented relatively underdeveloped libraries so that some primitive functions had to be written in 8088 Assembly Language. Moreover, because the Seequa Chameleon is a very slow computer, all video, graphics, and text functions are written in 8088 Assembly language to enhance the speed of execution.

The C programming language was chosen for the ECAT development primarily because of its capabilities for speed and flexibility in systems programming, and because it is by far the more popular computer language for microcomputer applications. ECAT is primarily written in HP-UX C, Release 5.0.1. All functions requiring high speed functions (especially graphics) were written in Motorola's 68000 Assembly.

The porting of software to a computer system other that the one for which it is designed is generally considered a difficult undertaking. It requires a thorough understanding of the software purpose as well as the knowledge of both computer systems. The fact that this porting effort required a change in the computer language, as well, made the porting effort even more difficult. The porting programmers depend upon (a) documentation of the original software and (b) the careful study of the operation of the original software.

It is reported by the ECAT developers that Project A hardware and software were not available at convenient times to make a side by side comparison of ECAT and Project A. In particular, ECAT programming efforts began in April 1990, at which time none of the Seequas made available to RGI by the Naval Personnel Research and Development Center (NPRDC) were in working

condition. Working hardware was made available to RGI programmers in September 1990; however, by that time much of the programming had been completed. Attempts were later made to do proper validation studies. These attempts are described later in this chapter.

Project A Computer Test Documentation. Documentation, from a software engineering standpoint, was not developed for the Project A software. Comprehensive descriptions of the source code, limitations, requirements, and pedestal use are not available. The Pascal code is not well-documented with explanatory text which, although Pascal purports to be a self-documenting language, was reported by the ECAT developers to have created considerable ambiguity for the porting of the software to the C language on the HP-IPC. ECAT documentation was prepared to be considerably more extensive including embedded comments in the software source code, formal project reports, and user manuals.

Stimulus Files. The Project A software used "stimulus files" to control the presentation of tests. These are files containing commands that are interpreted by the Project A test software. There is one stimulus file for each test and it determines the sequence of events, display of instructions, and the characteristics of items to administer. With these stimulus files (which were prepared by participating psychologists), various test and item parameters are manipulated to accomplish desired test item characteristics. The ECAT software uses the same stimulus files with some modifications in the text of instructions.

In addition to the stimulus file, one of the tests converted from Project A to ECAT (Target Identification) also used a data file. The additional data file contained the data used to construct the graphics representations of stimuli such as tanks and aircraft. ECAT used exactly the same format as Project A but some aesthetic changes were made. These changes were implemented because of differences in the appearance of the two kinds of graphical displays.

Use of Project A Tests in Porting. It is reported by RGI programmers that a Project A pedestal and demo software were only briefly available for study during ECAT's early design phases. It is said that the hardware eventually broke down and no other equipment was available during ECAT's actual program development and system testing phases, and that it was not possible to actually compare ECAT and Project A.

Response Pedestal Calibration. The range of Project A's joysticks varies across devices and computers. Because of this, the software was designed to use a linear adjustment constant for each device which was to have been pre-calibrated for each

computer/pedestal combination. This calibration step was not considered necessary in the ECAT implementation.

Functional Specifications. Specifications for Project A and ECAT were limited primarily to hardware and technical specifications. Software specifications were limited to descriptions of the overall purpose for each subtest. The most important technical specifications include: a computer clock resolution of 1 millisecond accuracy, a minimum joystick range of 110, and high resolution graphics.

Accelerated CAT-ASVAB (ACAP) Considerations. Originally, sponsors of ECAT considered merging ECAT with ACAP. The intention was to administer both testing systems in sequence (i.e., ACAP, followed by ECAT). Unfortunately, ACAP evolved into a large software module which drained system resources. This made it impractical to run ECAT and ACAP together. To date, ACAP and ECAT have never been administered together. Current plans call for enhancements of the HP-IPCs to enable them to run the two testing packages together. By December 1992, the upgrades are expected to be in place.

Additional (Navy) Specifications. Detailed functional specifications for the ECAT software were developed. These specifications included detailed hardware and computer requirements, protocols for interfacing computers with the pedestals, descriptions of the stimulus files, suggested algorithms for important aspects of software, equations for scoring and calculating paths and crosshair motions, detailed descriptions of the subtests, including story boards, and file layouts.

For the ECAT implementation, as already described, the menu dial was removed from the response pedestal, and a HELP key was added. ECAT was extended to include help sequences for: bad user input, instruction, practice and test timeouts, and explicit calls for help by pressing the HELP key. The layouts of the response files were changed in format and by including additional data.

Failure/Recovery. Failure/Recovery was implemented in both the Project A and ECAT software so that an examinee could restart at the beginning of the test where he or she was interrupted.

Pedestal Differences

Regardless of detailed specifications, differences or enhancements, sponsors of ECAT dictated that it should "work" essentially the same as Project A. The most critical factor in developing an analogous system was not in computer characteristics, programming languages, or graphics qualities, but in the "response pedestal" quality and functionality. This

section discusses physical and functional differences between the Project A and ECAT pedestals. The advantages and consequences of the differences are considered.

HELP Key (Addition of). The ECAT pedestal includes a HELP key. The size, shape and color of the key is identical to the HP-IPC keyboard function keys. The key is located at the upper left of the pedestal, approximately in the same place as the HELP key on both the ECAT and ACAP keyboards. The HELP key may be pressed at any time during test administration (including during the instructions and practice) to initiate a standard help sequence. The HELP key and help sequences mark distinct deviations from the Project A pedestal and software which have no examinee help features. The help features also make ECAT more similar to the operational ACAP system.

Menu Dial (Removal of). Project A's dial was used in the test battery for examinees to enter their social security numbers and respond by menu selection to several biographical items. The dial was not used in any of the actual tests. For the purposes of the Navy, it was decided that the dial had no significant and irreplaceable functions, so it was not included for the ECAT. Since the modified ECAT keyboard is already attached to the computer, the numeric keypad could more easily be used to enter personal data.

The joystick units for both pedestals are constructed of different materials. The Project A joystick has plastic component parts which wear quickly, and cause it to fall out of calibration rapidly. The ECAT joystick has metal component parts making it more durable and reliable. Finally, the ECAT pedestal is less subject to joystick calibration problems.

Joystick Range of Motion. The GAME port joysticks and slides of the Project A response pedestal provide an average range of 4 to 140 in the Seequa Chameleon. The actual analog to digital conversion involves the proprietary use of the CPU for counting so that, in porting to new computers, the values depend upon the speed of the CPU. The joysticks and slides of the RGI pedestal yield readings of approximately 4 to 245. The analog to digital conversion is performed by the RGI pedestal device so that digital information is provided directly to the software.

Buttons. On both pedestals, the buttons work in the same manner: a single byte contains the bit status of each button. If the bit is 1, the button is down. If the bit is 0, the button is up. Eight buttons can be accommodated: BLUE, YELLOW, WHITE, LEFT RED, RIGHT RED, LEFT HOME, RIGHT HOME, and HELP on the ECAT pedestal (one bit, the "HELP" bit, is unused in Project A).

HOME POSITION. The HOME POSITION is actually two buttons. It is defined as the combination of either the two GREEN buttons

on the left side or the two GREEN buttons on the right side being pressed simultaneously. There is a LEFT HOME and a RIGHT HOME POSITION. The Project A pedestal is actually wired so that both green buttons on a side must be depressed in order for the computer to detect the "home" button, whereas for purposes of future flexibility, the ECAT pedestal was designed to permit detection of all four of the green buttons separately.

Serial Numbers. Both pedestals have serial numbers displayed on a metal label on the top centered between the joysticks. In addition to this, the HPIB pedestals each have their unique serial number stored in their circuity (via dip switches). The serial number is stored in two bytes and is part of the 10 bytes sent when data is read from the hardware. The fact that the serial number can be obtained directly from the pedestal is an advantage for the ECAT pedestal. The Project A pedestal required operator identification of the pedestal for purposes of calibration of individual computer/pedestal combinations.

Potential Test Differences Arising from Software

The evolution of the Project A software occurred over the earlier years of the project as particular needs were determined on the basis of psychological expertise. Thus, it tended to have been developed around the limitations and strengths for the specific machine, operating system, and response pedestal that were available. Consequently, its implementation was lacking in terms of portability to other computer environments. On the other hand, the ECAT software was implemented under the premise that it may eventually be transported to yet another system.

Restating what has been discussed so far, the two sets of software were developed in under different conditions by persons with differing areas of expertise, for two different types of computers, and in two different computer languages. However, the scope and extent of efforts to maintain construct or psychometric equivalence between the original and the ECAT software is not clear.

One step made by the ECAT developers was to use the original stimulus files and emulate the same purposes for the stimulus file commands. Recall that the stimulus files consist of a command system for defining relevant properties of a subtest. This includes the text used for test-taking instructions as well as for presentation of items. The stimulus files for ECAT were nearly exactly the same ones used by Project A. Note that some changes were made to modify or improve the instructions.

An issue in the porting of the software to the HP-IPC involved the differences in the properties of the graphical displays. Because of the differences in resolution and aspect

ratio between the displays, it was necessary to rescale all coordinates used in presentation of graphical figures. However, the same figures were used with the needed modifications to emulate the original figures used in the Project A tests.

Ported Software Testing

Once a working version of ECAT was completed, the software underwent standard ALPHA and BETA testing. In response to concerns about the equivalence of the ECAT pedestal, various systems testing strategies were devised and implemented to quantify performance similarities and differences between the two pedestals. The following subsections describe actions taken to validate ECAT.

Systems Testing. ECAT was subjected to in-house (ALPHA testing) before its delivery. While vigorously testing ECAT, RGI found there was no real way to compare it to Project A due to the unavailability of Project A hardware and software. This is an unfortunate omission, since a significant point of the ECAT project was to convert and administer Project A tests on a new machine.

To minimize the room for error, ECAT programmers emulated exactly many of the designs and algorithms in Project A. This was done not only to use the original data and stimulus files, but also to lend credibility to a realistic conversion.

Field Testing. BETA testing included administering ECAT to Navy recruits locally (in San Diego, California). This testing provided an opportunity to refine and debug ECAT before using it for formal data collection studies.

In addition to training Test Administrators (TA's) to administer ECAT, RGI also trained them to notice and record problems with ECAT. The TA's report how well ECAT examinees accept ECAT, how easily they understand and use ECAT, and any technical problems caused by software or hardware bugs, or by design flaws. As an example, field testing showed that the early HPIB joystick stems were too thin and fragile. Thicker stems are now used. Systems testing by RGI and NPRDC shows that, even though the thicker stems yield a smaller range of motion, this has no functional effect on the software.

NPRDC initiated a study to definitively validate ECAT against Project A, and compare functionality and scoring of ECAT and Project A side-by-side. The study involved a test-retest-retest-retest design, alternating between Project A on the Seequa, and ECAT on the Integral. Both testing systems were modified to collect detailed data from the Tracking tests. RGI developed software to facilitate quantitative and qualitative

analysis of the tracking data collected. Information on the results of this study has not yet been published.

Summary

Two versions of the tests have been developed: the Project A version and the ECAT version. It was intended that both versions measure the same psychological constructs with the same psychometric properties.

Unfortunately, the circumstances of the two development efforts differed substantially. One important difference is that the ECAT effort was not anticipated, or even conceived, during the development of the Project A tests so that specifications for duplication of function were not prepared. This fact led to a difficult task for the ECAT developers in that they were required to prepare comparable tests in a new computer language and on a different computer with no more than an operating version of the Project A software, the original Pascal source files, and the supporting files for stimulus definition.

The microcomputers used for the two versions of the tests have substantially different physical display characteristics. Both have comparable graphical functions but substantially different degrees of resolution.

The response pedestal (response acquisition device) that was designed for the original test software remained very similar in the newer version of the tests. One potentially important difference is that the newer pedestal is only 19 inches wide, whereas the original is 20 inches wide, thus altering the difficulty, on the part of the examinee, of reaching the centrally located buttons from the "home" position. Another difference is the addition of a "HELP" button on the newer pedestal permitting interruption of the test by the examinee.

Fortunately, the functional properties of the original tests with respect to psychological content were contained in plain text files called stimulus files. These files, prepared by participating psychologists in the Project A research team, were used to control the item properties of test presentation. In using these same files for the newer ECAT versions of the tests, and with the care taken to duplicate the functions of stimulus file entries, the two versions of the tests appear likely to have the same content with respect to the psychological constructs measured. However, the potential for some differences in psychometric properties (such as item difficulty) seems substantial.

Given the differences in the type of developmental effort and the differences in the computer hardware used in the two test development efforts, it is quite unreasonable to expect precise

equivalence between the two batteries of tests. At this point, the degree of difference is clearly an empirical issue.

APPENDIX B

COLLECTION AND MANAGEMENT OF DATA

This appendix details the data collection procedures that were employed and the data base management techniques that were used during the course of this project. Data collection was performed primarily by RGI Incorporated with the American Institutes for Research (AIR) being responsible for one data collection at Mayport Naval Base. AIR was responsible for building, maintaining, and documenting the data base which incorporated the data collected for this project with existing Department of the Army data.

Data Collection Procedures

RGI Incorporated (RGI) collected data at Forts Benning, Knox, Jackson, and Leonard Wood at various time periods between June 1990 and July 1991. Although small differences occurred at each location, RGI implemented similar procedures at all sites for coordinating the testing, and collecting the data.

General Testing Procedures

This section presents the general testing procedures that RGI test administrators (TAs) implemented at all sites during data collection. Subsequent sections present procedures peculiar to each individual test site, and the results from the data collection procedures.

Except for the Fort Jackson and the 1991 Fort Benning data collections, RGI recruited, selected, and trained TAs locally. Training lasted two or three days depending on the complexity of the tests being administered and included at least one session with actual test supervision to monitor the effectiveness of the locally hired TAs. The Army Research Institute (ARI) trained the 1991 Fort Benning TAs and assisted them during the first test sessions. The TA at Fort Jackson was also trained by ARI. However, for that data collection, training was accomplished remotely by telephone and through explanatory scripts written specifically for the TA.

Each site's Reception Station provided the TAs with rosters of available troops as well as a schedule of when they would be available for testing during their first four days of processing. Only soldiers with no prior military service were tested. The TAs coordinated test session start times with the Reception Station personnel. Drill Instructors delivered the troops to be tested and picked them up after testing. The TAs arrived 30 to 60 minutes prior to testing to prepare the necessary paperwork and materials.

When the soldiers arrived for testing, the TAs sat them at their stations, delivered the verbal instructions, assisted them in completing the background information form (when it was collected), answered any questions, and started them on the test. During testing, the TAs monitored the examinees' progress, and made session notes of any problems that arose during the session, such as distractions or equipment failures.

After everyone finished the test, the TAs thanked the soldiers for their participation and excused them. The Drill Instructor either instructed the soldiers where to go at the completion of testing, or picked them up personally when they were through.

Overview of Data Collection Efforts

Data were gathered at four sites during the base period of this project. Following are descriptions of the data collected at each site which include the instruments administered, the order in which the instruments were administered, and the number of examinees who were tested. Table B-1 reports the test sites where each data collection took place, the time period during which the data were collected, the number of examinees, and the number of test sessions.

Table B-1

Data Collected at the Army Sites

Test/Study	Test Sites	Collection Dates	Sample Size	# of Sessions
Computerized Tests	Benning	6/15/90 - 6/21/90	437	10
ASP Faking				
A. Honest	Benning	6/01/91 - 7/31/91	340	3
B. Fake Good	Benning	6/01/91 - 7/31/91	274	3
C. Coached/Warned	Benning	6/01/91 - 7/31/91	228	3
D. Coached to Look Good	Benning	6/01/91 - 7/31/91	336	4
Practice Effects				
A. Target Identification	Knox	7/23/90 - 8/10/90	116	4
B. Two-Hand Tracking	Knox	7/23/90 - 8/10/90	113	4
ABLE Faking & Order Effects*				
A. Combination 1	Knox	7/23/90 - 8/10/90	209	7
B. Combination 2	Knox	7/23/90 - 8/10/90	206	8
C. Combination 3	Knox	7/23/90 - 8/10/90	200	7
D. Combination 4	Knox	7/23/90 - 8/10/90	205	8
ABLE Faking				
A. Fake Average w/Examples	Jackson	1/04/91 - 1/12/91	358	2
B. Fake Good w/Examples	Jackson	1/04/91 - 1/12/91	349	2
C. Fake Good	Jackson	1/04/91 - 1/12/91	344	3
D. Honest	Jackson	1/04/91 - 1/12/91	114	1
Biodata/ABLE	Leonard Wood	3/02/91 - 5/25/91	1,003	15
Assembling Objects/Biodata	Leonard Wood	3/02/91 - 5/25/91	1,561	16

*Combinations refer to the ABLE test-taking strategies and psychomotor/perceptual speed test administration orders described in Table B-2.

Fort Benning

Computerized Tests. The instruments for this data collection consisted of seven tests from the Project A battery that RGI had programmed to run on the ECAT hardware and software. (See Appendix A for a description the computer programming and apparatus developed for the ECAT project.) The first was a 199 item non-cognitive self-report instrument, the Assessment of Background and Life Experiences (ABLE). The remainder were the six Army tests in the ECAT battery. Three of these tested spatial abilities and had been given only in the paper-and-pencil medium in Project A: Assembling Objects, Spatial Orientation, and Spatial Reasoning. In this new computerized format, examinees responded to these tests by pressing keys on a standard computer/typewriter keyboard. For the final three tests, examinees responded on a custom-built response pedestal with buttons, sliding controls, and joysticks. These three tests were of psychomotor abilities (One-Hand and Two-Hand Tracking) and perceptual speed and accuracy (Target Identification). RGI provided the computers and software for testing. The Navy Personnel Research and Development Center (NPRDC), which is the developer of ECAT, supported this research by permitting the use of the ECAT hardware.

Data for this research were collected between 15 June and 21 June 1990. Depending on the number of soldiers arriving at the Reception Station, one or two test sessions were conducted each day. Because of space limitations, testing was limited to no more than 48 examinees at one time. The test sessions lasted approximately 2.5 - 3 hours.

ASP Faking. During June and July 1991, data were collected using the Adaptability Screening Profile (ASP). This test consisted of two parts: Part I was a shortened version of the Armed Services Adaptability Profile (ASAP) and Part II was a shortened version of the ABLE. One to three test sessions consisting of up to 125 soldiers each were conducted on Saturdays to obtain the analysis population.

The test was administered under four different conditions. There were a total of 1,178 male soldiers tested. Table B-1 details the number of soldiers in each condition.

Fort Knox

Between 23 July and 10 August 1990, data were collected for two practice effect studies (one for Two-Hand Tracking, the other for Target Identification), an ABLE faking study and an order effects study on the two tracking tests and Target Identification. One to four test sessions a day were conducted between the hours of 7:30 am and 5:30 pm, Monday through Friday. The number of and size of test sessions were determined by the

number of recruits arriving at the Reception Station, and the scheduling of normal processing requirements. A maximum of 33 soldiers could be tested at one time. The data were collected separately for the two practice effects studies, but the collections for the ABLE faking study and the order effects study were combined. Soldiers were included for testing in either (1) one of the two practice effects studies, or (2) the combined ABLE Faking Study/Order Effects studies. Collections for the studies were rotated by test session to control for possible effects of starting times on test performance. The TAs requested the soldiers not to discuss the test with anyone during their four days at the Reception Station to prevent possible contamination of future test takers.

Practice Effects. The studies required a minimum of 100 soldiers tested on each of the Target Identification and Two-Hand Tracking test sequences. To meet this requirement, 20 to 33 soldiers were tested per session across four sessions each for Target Identification and Two-Hand Tracking. Soldiers participated in only one of the two studies. Table B-1 details the final counts.

ABLE Faking and Order Effects. Soldiers participating in these studies were administered the three psychomotor/perceptual speed tests (One-Hand Tracking, Two-Hand Tracking, and Target Identification), followed by the ABLE. Each of four different test administration orders for the psychomotor/perceptual speed tests were paired with one of four different ABLE test taking strategy instructions. Table B-2 shows the combinations of ABLE test taking strategies and psychomotor/perceptual speed test sequences. For example, all soldiers who were administered the ABLE "Honest" test taking strategy also completed the sequence of One-Hand Tracking, followed by Two-Hand Tracking, and Target Identification.

A minimum of 200 soldiers were needed for each of the four combinations for a minimum total of 800 examinees. To meet these requirements, the TAs tested between 7 and 33 soldiers per session across 30 test sessions. Table B-1 details the number of soldiers per session and the number of sessions.

Table B-2

ABLE Test Taking Strategy and Psychomotor/Perceptual Speed Test Order Combinations

	ABLE Test Taking Strategy	Psychomotor/Perceptual Speed Test Order
Combination 1	Honest	One-Hand Tracking, Two-Hand Tracking, Target ID
Combination 2	Fake Good, No Examples	Two-Hand Tracking, One-Hand Tracking, Target ID
Combination 3	Fake Good, With Examples	One-Hand Tracking, Target ID, Two-Hand Tracking
Combination 4	Fake Average, With Examples	Two-Hand Tracking, Target ID, One-Hand Tracking

Fort Jackson

Between 4 January and 12 January 1991, the ABLE was administered at Fort Jackson as a continuation of the ABLE Faking Study conducted at other Army posts. The final count, as detailed in Table B-1, was 1,165 examinees of whom approximately 45 percent were female.

Fort Leonard Wood

Data were collected at Fort Leonard Wood between 2 March and 25 May 1991. The data were intended to support two studies and included three tests: a 130-item version of the ABLE, a 120-item Biodata Questionnaire, and the Assembling Objects test. The two studies were a comparison of the ABLE and Biodata Questionnaire, and an examination of existing items in the Assembling Objects versus newly designed potential items.

The Assembling Objects test was administered in 15 versions with each version having 48 total items. In versions 1 through 12, there were the 36 items from previous administrations of the test, and 12 new items per version that were being administered for possible inclusion in an updated version of the test. Versions 13 through 15 included only the new items. Over the 15 versions of the test, there were 144 new items. The 36 existing items were the same in versions 1 through 12, but the new items were distributed among the versions with each version including 12 unique new items. Version 13 included the new items from versions 1 through 4; version 14 included the new items from versions 5 through 8; and version 15 included the new items from versions 9 through 12.

Data for both studies were collected during the soldiers' second or third day of processing, before being shipped to basic training. Depending on the number of soldiers arriving and the requirements for processing, TAs tested the soldiers on most Saturdays after breakfast and, when necessary, Tuesday through Thursday after dinner. Normally, the TAs administered two test sessions a week.

The ABLE and the Biodata Questionnaire were combined in one test booklet. This test was administered to the first 1,000 soldiers, who were expected to complete the test in a maximum of 90 minutes. The Assembling Objects test was administered consecutively with the Biodata Questionnaire to the next 1,500 soldiers, and these tests were expected to last a maximum of 52 and 50 minutes respectively.

During the administration of these tests, misprints were found in several of the ABLE items. These misprints were remedied with specific directions to examinees. They were also brought to the attention of ARI in order to make corrections in

future test administrations, and to alert analysts of possible complications in the interpretation of analysis results.

Mayport

On 3 March 1991 the complete predictor battery was administered to four sailors at the Mayport Naval Base. The tests were given in a mix of computer administered and paper-and-pencil tests. The computer administered tests included: Simple Reaction Time, Choice Reaction Time, Short Term Memory, One-Hand Tracking, Two-Hand Tracking, Perceptual Speed and Accuracy, Number Memory, Cannon Shoot, Target Identification, and Target Shoot. The paper-and-pencil tests included: Assembling Objects, Map, Spatial Reasoning, Spatial Orientation, Mazes, and Object Rotation.

These data were processed in the same manner as all of the other CMOS data. The raw "item-level" data, i.e. the responses provided by the sailors, were uploaded to the NIH computer system. The item-level responses were scored by individual test according to the rules devised under Project A/Career Force. Then the individual test scores were combined into composite scores. As this work was carried out under a Military Interdepartmental Purchase Request from the Naval Air Test Center, Patuxant River, its results were reported to that command for analysis and reporting.

Integrated Research Data Base Management

The CMOS project was designed to implement and upgrade testbeds of measures that had been developed under Project A/Career Force in certain critical MOS. The data collected included various types and combinations of predictor and criterion data. The data collected for each component of CMOS were determined in part by the goals of the particular analysis and in part by the availability of subjects. Usually the data collected for this project were identified by the site at which they were collected. Similar data were collected at multiple sites, but were rarely combined for any studies. Data collections were designed primarily to fit the requirements of a specific component.

Although data from multiple sites were not generally combined for analysis purposes, analysis data were not limited to the testing data collected on site. The data were often combined with demographic data, applicant/accessions data, and the Enlisted Master File to provide additional analysis variables.

Data collected under the aegis of the CMOS project are in many ways a continuation of the data collected under Project A/Career Force research. Data set and variable naming conventions were developed under Project A that identify at a

glance the type of data, the location where it was collected and the time period during which it was collected. These rules were rigidly enforced for Project A/Career Force and provided for ease of analysis over more than 10 years and many changes in personnel.

Because the data collected for CMOS was an extension of the Project A/Career Force measures and the naming conventions were familiar to the researchers, many of whom also worked on Project A/Career Force, similar data set and variable naming conventions were followed under CMOS. Certain modifications had to be made, however. The Project A/Career Force data were collected as three validation studies. Massive data collection efforts included multiple sites, took months to complete, and were designed to support all forms of analysis. Under CMOS, the data were collected at single sites, usually over shorter time spans, for a particular study. These differences necessitated changes primarily in the data set naming conventions. Similar security precautions used in Project A/Career Force were also followed for the CMOS data sets (Wise, Wang & Rossmeissl, 1983).

In addition, data were archived in this project data base from two testbeds that started before CMOS: the Skills Selection and Sustainment Program (S3) of the U.S. Army Training and Doctrine Command (TRADOC) (Walker, 1989), and an evaluation of gunnery aptitudes in the deployed Army forces in Germany. These two testbeds included measures that had not been collected under Project A/Career Force, primarily of performance on real and simulated weapons systems.

The data collected under the CMOS project have been amalgamated into the CMOS Integrated Research Data Base (IRDB). The IRDB includes all of the individual data sets gathered under the auspices of the project.

Overview Of IRDB Contents

The IRDB contains data from a variety of sources and sites. Some of the data, such as scores on the Armed Services Vocational Aptitude Battery (ASVAB) and identifiers of individuals' Military Occupational Specialties (MOS), were obtained from military personnel data bases. Other data were obtained from ARI archives on testbeds that started before this project. Finally, there were original data generated by original research under the present project.

Two methods were used to record the data. Some of the data were obtained using paper-and-pencil tests where the responses were recorded by filling in a scannable answer sheet. Other data were obtained using instruments administered through Enhanced Computerized Assisted Testing (ECAT) where the responses were entered directly to a diskette.

The IRDB consists of three different kinds of data system files in order to accommodate the different users of the files. The different kinds of data file are:

- separate files for each site/data collection containing person-level data on all instruments administered during a unique implementation,
- separate files containing item-level data for individual instruments, and
- integrated "summary" files containing scores from several or all of the instruments included in the data set.

These will be discussed more fully below.

Contents of the IRBD

The IRDB contains multiple types of data because of the varied sources and requirements for testing at each site. The sources of data included in the IRDB are:

Basic Demographic Information (e.g., gender, race)

Applicant/Accession Information (e.g., ASVAB scores)

Experimental Predictor Measures

- Biodata/Temperament Measures (ABLE)
- Spatial Tests
- Psychomotor Tests

Operational Sources

- Enlisted Master File

Naming Conventions

Data Set Names. Rules were established for the naming of the individual sets which comprise the IRDB. Initially, separate data sets were created for each data collection site. These files may combine multiple instrument types within the same data set, and are organized at the level of the individual soldier. Between sites, variable names (to be discussed below) use a single set of naming conventions for ease of combining across sites. From these initial data sets, instrument specific data sets can be created that combine unique instrument information from multiple sites.

The names of data sets containing site specific implementation data consist of:

1. A one letter C-MOS project identifier to be used on all data sets, 'C',

2. the letter 'S' to indicate it is site-specific,

3. a two digit site identifier,

4. a one character sequencing identifier (A-Z, 1-9) to indicate the order in which the test was administered, and

5. a two character version identifier.

The two digit site identifier was drawn from the following list:

01	Fort Benning
02	Fort Bliss
07	Fort Knox
14	USAREUR
21	Fort McClellan
22	Fort Jackson
23	Fort Leonard Wood

The two digit version number consists of a 'V' and one number starting with 1 and continuing as needed. The initial reading of the raw data was V1; V2 contains updates to the V1 data, etc. Using these naming conventions, the initial data set for the first test(s) administered at Fort Benning was named CS01AV1.

Certain analyses required specific instrument data collected from multiple sites. Therefore, in addition to the site specific data, instrument specific data sets were created combining data from multiple sites.

The names of data sets containing instrument specific data consisted of:

1. A one letter CMOS project identifier to be used on all data sets, 'C',

2. the letter 'I' to indicate it is instrument specific,

3. a two character instrument identifier, and

4. a three character sequencing identifier.

All instrument specific data sets begin with the letters 'CI' to indicate they contain CMOS instrument specific data. The two character instrument identifiers include[1]:

AB - ABLE Data
AO - Assembling Objects
AV - ASVAB Data
MZ - Mazes
OR - Object Rotation
OT - Orientation
RS - Spatial Reasoning
T1 - One-Hand Tracking
T2 - Two-Hand Tracking
TI - Target Identification

The three digit version number consists of a 'V' and two numbers starting with 01 and continuing as needed. The initial reading of the raw data was V01, V02 contains updates to the V01 data, etc. Using these naming conventions, the first data set of ABLE data was named CIABV01.

Variable Names. A set of rules regarding the naming of variables was devised in order to minimize confusion and to provide as much information as possible to future researchers. The naming convention for instrument data as initially collected by site combined:

- a two to four character prefix indicating data source, and
- up to six characters to identify specific variable attributes.

Note that eight characters is the maximum variable name length in most statistical packages available for analyses of the data.

The two to four characters indicating the data source include the following codes[2]:

---

[1] This list represents the full range of possible instrument-specific data sets. At the time this report was written, not all identifiers had actually been used.

[2] This list represents the full range of possible data sources. At the time this report was written, not all of the possible data sources had been accessed.

AC - Accessions Data
EM - EMF Data
CH - DMDC Cohort
SQ - SQT Data
A087 - ABLE Data, 87 item version
A199 - ABLE Data, 199 item version
A202 - ABLE Data, 202 item version
A209 - ABLE Data, 209 item version
A291 - ABLE Data, 291 item version
AO - Assembling Objects
MZ - Mazes
OR - Object Rotation
OT - Orientation
RS - Spatial Reasoning
H1 - One-Hand Tracking
H2 - Two-Hand Tracking
TI - Target Identification

The ABLE data were collected using different versions, each with a different number of total variables in the instrument. For ease of identification, variable names for each version of the ABLE begin with a four character prefix indicating the total number of items in the test and continue to number the items from 001 through the largest number of items. For example, the items in the 87 item version were named A087I001 - A087I087, and the items in the 291 item version were named A291I001 - A291I291. A separate file was also maintained which indicated which items across the versions are equivalent in order that the same items can be pulled from the various versions if necessary.

Some variables appear across multiple data sets, but are neither site specific nor instrument specific. These variables, such as name and Social Security Number, were named using the following conventions:

- a two character prefix indicating project wide applicability, 'CM', and
- an up-to-six character item label.

When creating analysis variables for the use of researchers, similar variable naming conventions were used. Analysis variable names consist of:

- A two character prefix indicating data source, and
- a six character score label.

The two characters indicating the data source are the same as those already listed.

The six character score label indicates the type of variable. Score variables consist of mnemonic descriptors that were developed as needed.

Data Base Storage And Security

File Structure

As described above there are three different kinds of data system files included in the IRDB. The different file types are included in order to accommodate different uses of the information based on the analysis needs. The three types of files include:

- separate files for each site/data collection containing person-level data on all tests administered during a unique implementation,
- separate files containing item-level data for individual instruments, and
- integrated "summary" files containing scores from several or all of the instruments included in the data set.

The intention in designing this system of data files was to provide maximal efficiency in generating workfiles for analysis. A compromise was achieved between a storage-efficient system of independent "relations" where each data element was stored only once (but where a great deal of "relating" is necessary to create required workfiles), and an overly integrated system where all of the information on all soldiers was stored in a single file (even though many elements did not apply to some soldiers).

Processing Procedures

Our general approach to processing data collected in the field involved the following steps:

1. Log in and check the completed instruments.
2. Convert scannable data to machine readable form.
3. Edit individual data files, and process missing items within each instrument.
4. Edit for consistency and completeness across instruments.
5. Merge in operational and previously collected data as appropriate.
6. Compute scores needed for analyses.
7. Generate workfiles as required for analysis.
8. Add new summary variables as they were generated during analyses.

9. Create final archiving information at the completion of the project.

Data have been collected for this project using two types of instruments:

1. Computer-administered tests.
2. Paper-and-pencil tests.

In the following sections, the general approach to processing data from these instruments will be described.

Computer-Administered Tests. Data from these tests were stored in diskette form and needed to be re-processed before being uploaded to NIH. The following system was used for processing the computer data.

Data Upload. Data collected on diskette were uploaded to the NIH mainframe using KERMIT. This system worked well for this project despite KERMIT's relatively slow processing speed, because the amount of data collected in this fashion was small.

Initial Checking. Data were copied onto an NIH tape as a backup, and also onto disk for initial processing and checking. The initial checking included editing out-of-range values, checking for duplicate records within each test site, and analyzing missing values.

Data Scoring. Existing programs from Project A/Career Force were adapted for scoring the computer battery as required. These programs created one or more scores for each test in the battery. In addition, new programs were written where they were required by the analyses.

Merge Data. Data from background information along with other existing Army data were added to the files as needed.

Analysis Files. Data were prepared and turned over to the analysis team for data analysis.

Paper-and-Pencil Tests. For instruments using scannable answer sheets, the answer sheets were shipped to RGI for scanning. Each answer sheet was checked to ensure maximum accuracy prior to being scanned. Sheets were examined to verify the SSN grid and to check for stray marks. After scanning, the following system was used for processing.

Initial Checking. Data were copied onto tape at NIH as a backup and also onto disk for initial processing and checking. The initial checking included editing out-of-

range values, checking for duplicate records within each test site, and analyzing missing values.

Data Scoring. Scoring programs were developed for scoring the spatial tests. Since initial scoring programs already existed from Project A/Career Force, they were retrieved and tailored to meet the needs of the analyses for this project. Additional programs were created as needed.

Merge Data. Data from background information along with other existing Army data were added to the file as needed.

Analysis Files. The data were prepared and turned over to the analysis team for data analysis.

Data Dissemination and Documentation

The data collected for the CMOS project were specifically designed to focus on particular instruments in order to provide detailed analyses to confirm earlier Project A/Career Force analysis findings and to enhance understanding of the instruments themselves. This resulted in the creation of a series of restricted data sets that were used primarily on a stand-alone basis. These individual data sets can be easily integrated in future analyses with a minimum of effort because they were created using consistent data set and variable naming conventions.

The data were provided to analysts in response to workfile requests submitted to the IRDB Manager who created individual analysis data sets according to the specifications of the analyses. When the data sets were created, documentation was provided to the analysts in the form of variable lists, definitions of variable values, and descriptions of any idiosyncracies found in the individual file.

Data Base File Directory

Although the IRDB plan was formulated to provide strict naming conventions, there have been data sets created under this project which do not conform entirely with those conventions. In addition, not all data sets created for this project have been under the centralized control of the IRDB Manager. Therefore, the following list of data sets currently available contains data sets whose names do not conform to all conventions. It is not an exhaustive list of data available for analysis, it contains only those data sets created by the IRDB Manager. The data sets and their creation dates are as follows:

Creation
Date Data Set Name

01/16/92 DSN=WTFJDHJ.SAS.LABLESC1
DESCRIPTION: Sample of Career Force LV examinees in infantry MOS with paper & pencil predictor data and who have AFQT scores.

12/05/91 DSN=WTFJDHJ.SAS.CISPAT01
DESCRIPTION: Assembling Objects, Orientation, Reasoning data collected at Benning in June 1990 using the computerized versions of the tests.

12/02/91 DSN=WTF1ARE.SAS.CIAOV02
DESCRIPTION: Workfile of Leonard Wood Assembling Objects data for H. Busciglio.

11/22/91 DSN=WTFJDHJ.SAS.CS01D1A6
DESCRIPTION: Rescore spatial tests collected at Benning in June 1990 using Project A/Career Force rules.

11/18/91 DSN=WTFJDHJ.SAS.CS01D1A5
DESCRIPTION: Score the ABLE data collected at Benning in June 1990, and correct scoring errors in the spatial tests.

09/18/91 DSN=WTFJDHJ.SAS.CS14AV4
DESCRIPTION: Attempt to add ASVAB data from Sep 90, Dec 90, and March 91 EMF files.

09/17/91 DSN=WTFJDHJ.SAS.CS23BV2
DESCRIPTION: Create individual item level data for the 36 "old" Assembling Objects items, and the 144 "new" items collected at Leonard Wood.

09/12/91 DSN=WTF1ARE.SAS.CIABV03B
DESCRIPTION: Workfile of ABLE scale scores and factor scores for data collected at Benning in June 1990 using the computerized version of the ABLE test for Jay Silva.

09/10/91 DSN=WTF1ARE.SAS.CIABV03
DESCRIPTION: Workfile of ABLE item level responses and response latencies collected at Fort Benning in June 1990 using the computerized version of the ABLE test for Jay Silva.

09/09/91 DSN=WTFJDHJ.SAS.CS14AV3
DESCRIPTION: Add a numeric version of the Table VIII variable to the Germany Gunnery Data.

09/07/91 DSN=WTFJDHJ.SAS.CS14AV2
DESCRIPTION: Add ASVAB and AFQT from the EMF to the Germany Gunnery data and score the Mazes and Orientation Tests.

08/26/91 DSN=WTF1ARE.SAS.CIABV02
DESCRIPTION: Workfile of 130 item ABLE data collected at Ft. Leonard Wood for Len White.

08/26/91 DSN=WTF1ARE.SAS.CIAOV01
DESCRIPTION: Workfile of the new Assembling Objects data collected at Ft. Leonard Wood. DO NOT USE THIS DATA SET. THE VARIABLES HAVE NOT BEEN DEFINED CORRECTLY. USE WTF1ARE.SAS.CIAOV02.

08/26/91 DSN=WTF1ARE.SAS.CIBIOV2
DESCRIPTION: Workfile of the 130 ABLE items and the 120 biodata items collected at Ft. Leonard Wood for Fred Mael.

08/26/91 DSN=WTF1ARE.SAS.CIBIOV1
DESCRIPTION: Workfile of the new biodata only collected at Ft. Leonard Wood for Fred Mael.

08/26/91 DSN=WTFJDHJ.SAS.CS23AV1
DESCRIPTION: 130 item ABLE data and 120 biodata items collected at Ft. Leonard Wood in February through April 1991.

08/26/91 DSN=WTFJDHJ.SAS.CS23BV1
DESCRIPTION: New spatial (Assembling Objects) and biodata collected at Ft. Leonard Wood in February through April 1991.

07/08/91 DSN=WTF1ARE.SAS.CIAVB01
DESCRIPTION: Combine the Jackson and Knox ABLE Faking Study data for Len White and Mark Young.

06/20/91 DSN=WTFJDHJ.SAS.CS22V03
DESCRIPTION: Create ABLE scale score variables and factor score variables using rules devised in Career Force for Jackson ABLE Faking Study data.

06/20/91 DSN=WTFJDHJ.SAS.CS22V02
DESCRIPTION: Score the Jackson ABLE Faking Study data according to the same rules used for the Knox ABLE Faking Study data.

06/18/91 DSN=WTFJDHJ.SAS.CS22V01
DESCRIPTION: Read in the raw data for the ABLE Faking Study collected at Fort Jackson in January 1991.

05/15/91 DSN=WTFJDHJ.SAS.CS07TIV3
DESCRIPTION: Calculate ASVAB factor scores in the Knox Target ID Practice Effects Data.

05/15/91 DSN=WTFJDHJ.SAS.CS07T2V3
DESCRIPTION: Calculate ASVAB factor scores in the Knox 2-Hand Tracking Practice Effects Data.

05/15/91 DSN=WTFJDHJ.SAS.CS07D5V2
DESCRIPTION: Calculate ASVAB factor scores in the Knox Order Effects data.

04/30/91 DSN=WTFJDHJ.SAS.CS14AV1
DESCRIPTION: Gunnery data collected in Germany in 1989. Includes Spatial Test Scores (Maze and Orientation), ABLE raw and percentile scores and an Overall Spatial/Tracking score. All score variables were calculated by the data collectors.

04/05/91 DSN=WTFJDHJ.SAS.CS07AV1
DESCRIPTION: Orientation, Maze, ABLE (187 items), 1-Hand, 2-Hand, data collected at Knox February 1987 - May 1989.

04/04/91 DSN=WTFJDHJ.SAS.CITIV01
DESCRIPTION: Target ID test results from the Knox Order Effects Study and the 1st sitting from the Knox Practice Effects study combined.

04/03/91 DSN=WTF1ARE.SAS.MAYPRTV5
DESCRIPTION: Workfile of Mayport data.

03/28/91 DSN=WTFJDHJ.SAS.MAYPRTV1
DESCRIPTION: Read in computer predictors collected on 4 sailors at Mayport March 1991.

03/28/91 DSN=WTFJDHJ.SAS.MAYPRTV2
DESCRIPTION: Read in paper & pencil predictors collected on 4 sailors at Mayport and add to computer predictors.

03/28/91 DSN=WTFJDHJ.SAS.MAYPRTV3
DESCRIPTION: "Clean" item level computer predictor data collected at Mayport.

03/28/91 DSN=WTFJDHJ.SAS.MAYPRTV4
DESCRIPTION: Compute basic scores on Mayport predictor data.

03/28/91 DSN=WTFJDHJ.SAS.MAYPRTV5
DESCRIPTION: Create composite scores for Mayport computer predictor data.

03/27/91 DSN=WTFJDHJ.SAS.CS07D1A4
DESCRIPTION: Score 1-Hand, 2-Hand, and Target ID in Benning Predictors using rules devised in Project A/Career Force.

03/20/91 DSN=WTFJDHJ.SAS.CS01D1A3
DESCRIPTION: Merge ASVAB scores from EMF with Benning Predictors.

03/19/91 DSN=WTFJDHJ.SAS.CS01D1A2
DESCRIPTION: Unsuccessful attempt to merge ASVAB scores from Accessions data with Benning Predictors.

03/12/91 DSN=WTFJDHJ.SAS.CS07D5A2
DESCRIPTION: Knox Order Effects data with ASVAB scores and scores calculated using rules devised in Project A/Career Force.
Order = 1-H, 2-H, TID

03/12/91 DSN=WTFJDHJ.SAS.CS07D5B2
DESCRIPTION: Knox Order Effects data with ASVAB scores and scores calculated using rules devised in Project A/Career Force.
Order = 2-H, 1-H, TID

03/12/91 DSN=WTFJDHJ.SAS.CS07D5C2
DESCRIPTION: Knox Order Effects data with ASVAB scores and scores calculated using rules devised in Project A/Career Force.
Order = 1-H, TID, 2-H

03/12/91 DSN=WTFJDHJ.SAS.CS07D5D2
DESCRIPTION: Knox Order Effects data with ASVAB scores and scores calculated using rules devised in Project A/Career Force.
Order = 2-H, TID, 1-H

03/12/91 DSN=WTFJDHJ.SAS.CS07D5V1
DESCRIPTION: Combine the 4 Order Effects data sets into one data set.

03/12/91 DSN=WTFJDHJ.SAS.CS07TIV2
DESCRIPTION: Knox TID Practice Effects data plus ASVAB scores and scores calculated using rules devised in Project A/Career Force.

03/06/91 DSN=WTFJDHJ.SAS.CS07T2V2
DESCRIPTION: Knox 2H Practice Effects Study data plus ASVAB scores and scores calculated using rules devised in Project A/Career Force.

02/20/91 DSN=WTFJDHJ.SAS.CS01D1A1
DESCRIPTION: Predictor data collected from Benning June 1990 (ABLE (199 items), Assembling Objects, Spatial Orientation, Spatial Reasoning, 1-Hand, Target ID, 2-Hand).

01/24/91 DSN=WTFJDHJ.SAS.CS07D5A1
DESCRIPTION: 1-Hand, 2-Hand, Target ID data collected at Knox July/August 1990 ("Order Effects Study"). Order = 1-H, 2-H, TID

01/24/91 DSN=WTFJDHJ.SAS.CS07D5B1
DESCRIPTION: 1-Hand, 2-Hand, Target ID data collected at Knox July/August 1990 ("Order Effects Study"). Order = 2-H, 1-H, TID

01/24/91 DSN=WTFJDHJ.SAS.CS07D5C1
DESCRIPTION: 1-Hand, 2-Hand, Target ID data collected at Knox July/August 1990 ("Order Effects Study"). Order = 1-H, TID, 2-H

01/24/91 DSN=WTFJDHJ.SAS.CS07D5D1
DESCRIPTION: 1-Hand, 2-Hand, Target ID data collected at Knox July/August 1990 ("Order Effects Study"). Order = 2-H, TID, 1-H

01/24/91 DSN=WTFJDHJ.SAS.CS07T2V1
DESCRIPTION: 2-Hand Tracking data collected at Knox July/August 1990 ("2H Practice Effects Study").

01/24/91 DSN=WTFJDHJ.SAS.CS07T1V1
DESCRIPTION: Target ID data collected Knox July/August 1990 ("TID Practice Effects Study").

01/02/91 DSN=WTFJDHJ.SAS.CS07BCV1
DESCRIPTION: ABLE (202 items) collected at Knox July/August 1990 ("ABLE Faking Study").

01/02/91 DSN=WTFJDHJ.SAS.CS07BCV2
DESCRIPTION: ABLE faking data scored at item level.

01/02/91 DSN=WTFJDHJ.SAS.CS07BCV3
DESCRIPTION: Create ABLE scale scores and factor scores using L. White rules.

12/14/90 DSN=WTFJDHJ.SAS.CS07SCR1
DESCRIPTION: ABLE item scores for Knox ABLE Faking Study.

12/12/90 DSN=WTFJDHJ.SAS.CLINK02
DESCRIPTION: Demographic data collected at Knox July/August 1990 added to CLINK01.

11/13/90 DSN=WTF1ARE.SAS.CS21ACV1
DESCRIPTION: Workfile for Len White.

11/07/90 DSN=WTFJDHJ.SAS.CLINK01
DESCRIPTION: Demographic data collected at McClellan July/August 1990.

11/07/90 DSN=WTFJDHJ.SAS.CS21ACV1
DESCRIPTION: ABLE (250 items) collected at McClellan July/August 1990.

07/11/90 DSN=WTFJDHJ.SAS.T1TACRV1
DESCRIPTION: Benning predictor (1-Hand, 2-Hand, ABLE (87 items), Orientation, Maze) merged with Gunnery data.

07/04/90 DSN=WTFJDHJ.SAS.T1CRITV1
DESCRIPTION: Gunnery data collected at Benning.

07/04/90 DSN=WTFJDHJ.SAS.T1CRITV2
DESCRIPTION: Gunnery data collected at Benning. This is the data set to use, as it contains more observations than the V1 data.

03/05/90 DSN=WTFJDHJ.SAS.CS02AV2
DESCRIPTION: Calculate means of items on tracking tests collected at Bliss, plus a standard overall score.

03/02/90 DSN=WTFJDHJ.SAS.CS02AV1
DESCRIPTION: Orientation, Maze, 1-Hand, 2-Hand data collected at Bliss in February-April 1988.